AFIT/GOR/ENS/99M-09

SELECTION OF PSYCHOPHYSIOLOGICAL FEATURES ACROSS SUBJECTS FOR CLASSIFYING WORKLOAD USING ARTIFICIAL NEURAL NETWORKS

THESIS

Trevor I. Laine
Captain, USAF

AFIT/GOR/ENS/99M-09

Approved for public release; distribution unlimited

AFIT/GOR/ENS/99M-09

SELECTION OF PSYCHOPHYSIOLOGICAL FEATURES ACROSS SUBJECTS FOR CLASSIFYING WORKLOAD USING ARTIFICIAL NEURAL NETWORKS

THESIS

Presented to the Faculty of the Graduate School of Engineering

Of the Air Force Institute of Technology

Air University

In the Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Operations Research

Trevor I. Laine, B.S., M.B.A.

Captain, USAF

March 1999

AFIT/GOR/ENS/99M-09

SELECTION OF PSYCHOPHYSIOLOGICAL FEATURES ACROSS SUBJECTS FOR CLASSIFYING WORKLOAD USING ARTIFICIAL NEURAL NETWORKS

Trevor I. Laine, B.S., M.B.A.
Captain, USAF

Approved:

Dr. Kenneth W. Bauer, Jr. (Advisor)	5 MAR 99 date
Dr. Jeffrey W. Lanning (Reader)	5 Mar 99 date
Dr. Glenn F. Wilson (Reader)	5 Mar 99 date

Preface

The focus of this thesis includes addressing three research objectives. These objectives are all related to the classification of aircrew workload in a simulated multi-task environment using psychophysiological input features. The first objective was to assess the differences between linear and non-linear models for both workload classification accuracy and the selection of the most important or *salient* features. The next objective was to determine if a single set containing the minimal, or *parsimonious*, number of salient input features for a group of subjects could be determined. If found, the third research question was to determine if a single artificial neural network (ANN) could be trained and used to accurately classify workload for all subjects. In other words, the research question "*Can one net fit all?*" was posed.

From the conception of the above research objectives to the final answers provided within this text, I owe a debt of gratitude to many people. First, I would like to thank my thesis advisor, Dr. Kenneth W. Bauer, and Capt Kelly A. Greene for introducing and solidifying my choice on this thesis topic through their genuine interest and observed energy in this research. Specifically, Dr. Bauer provided some of the essential tools for this research, including the derivation of a useful equation for determining discriminant loadings and the conception of the Signal-to-Noise (SNR) saliency measure. Additionally, I would like to thank Dr. Bauer for allowing me the latitude to focus this research in whatever direction was interesting to me.

I also thank Capt Greene for providing the background and support of this

research that stemmed from her Ph.D. dissertation where the first "real-world" application of the SNR feature screening was performed. I thank Capt Jeffrey W. Lanning for his continued interest in the research and his suggestions for incorporating temporal information into graphs that proved to be insightful. My appreciation also goes to Dr. Steven C. Gustafson for his presentation of physical models and insights, which sparked my interest and increased my understanding of pattern recognition.

I would also like to thank Dr. Glenn F. Wilson for his support of this effort, which included the opportunity to spend time in the laboratory, helping me to understand what "psychophysiological" features were and why they can be used as indicators of operator workload. My thanks also goes to Chris A. Russell for sharing some of his workload classification experience and insights using ANNs along with the code used to load and process the raw data. Additionally, I would like to thank the other scientists and engineers at the Air Force's Flight Psychophysiology Laboratory involved in collecting and providing the data used for this thesis effort. Finally, I am thankful for the funding provided by Dr. John F. Tangney of the Air Force Office of Scientific Research (AFOSR).

But, above all, I am indebted to my fiancée Lisa, who moved to Ohio and provided the understanding and support that was crucial to my completion of this research. In closing, I extend my thanks to family and friends for their support of my continued education and Air Force career that has kept me far from Oregon.

Trevor I. Laine

Table of Contents

List of Figures

List of Tables

Abstract

The issue of pilot workload is important to the United States Air Force because pilot overload or task saturation leads to decreases in mission effectiveness. Additionally, in the most extreme cases, pilot overload may lead to the loss of aircraft and crewmember lives. Current research efforts are utilizing psychophysiological data including electroencephalography (EEG), cardiac, eye-blink, and respiration measures in attempt to identify workload levels.

The primary focus of this effort is to determine if a single parsimonious set of psychophysiological features exists for accurately classifying workload levels between multiple test subjects. To accomplish this objective, the signal-to-noise (SNR) saliency measure is used to determine the usefulness of psychophysiological features in feedforward artificial neural networks (ANNs). The SNR saliency measure determines the saliency, or relative value, of a feature by comparing it to a feature of injected noise. For this effort, 36 psychophysiological features were derived from the data collected as each subject completed simulated crewmember tasks using the Multi-Attribute Task Battery developed by NASA. These tasks were randomly presented to the subjects in blocks with three distinct levels: low, medium, and an overload level in which subjects could not complete all tasks.

SELECTION OF PSYCHOPHYSIOLOGICAL FEATURES ACROSS SUBJECTS FOR CLASSIFYING MENTAL WORKLOAD USING ARTIFICIAL NEURAL NETWORKS

I. Introduction

1.1 Overview

This research contributes to the advancement of modeling mental workload in a multi-task environment. The primary goal of this thesis effort is to identify if a single *parsimonious* set of *salient* features exists for accurate classification of mental workload by multiple subjects. As used above *parsimony* refers to excessive frugality leading to a minimal number of features, while *saliency* refers to selecting those features with the strongest or most prominent predictive power for the classification problem. Input feature selection was selected as the primary research goal because the accuracy of any classification model is highly dependent on the quality of the input. This can be summarized by the statement "*garbage-in, garbage-out*," where no matter how carefully a model is determined; it is ultimately limited by the quality of the input.

For this research, psychophysiological features will be utilized by both multivariate discriminant models and artificial neural networks (ANNs) to classify observations into one of three mental workload levels. To determine if a single parsimonious set of features exists, saliency screening methods are employed by both multivariate discriminant and ANN models. The primary purpose of utilizing the discriminant models is to provide a benchmark for classification accuracy and to identify a set of features that appear to be linearly salient. In

addition to identifying a parsimonious set of features for use by an ANN, this optimal set of model inputs will be used to answer the proposed research question, *"Can one net fit all?"*

1.2 Background

The human operator is a crucial component of modern Air Force systems. Increasing technological complexity and the resulting potential for cognitive overload may mandate the need for monitoring the operator's state [1]. Specifically, today's Air Force jobs such as air traffic control and the piloting of aircraft require complex cognitive processing to complete the multiple tasks required for mission accomplishment. Unfortunately, as performance demands increase beyond a threshold, operators may experience a condition of "overload" in which all tasks can not be accomplished. As a worst case example, between 1986 and 1995, the USAF lost 14 fighter pilots to G-induced loss of consciousness, of which 13 occurred during demanding portions of flight, with associated high mental workload conditions [2].

Multivariate methods of analysis, including the use of ANN models, can be used to analyze physiological data of an operator in the attempt to gain insight of the current mental workload level. If accurate, this insight may be used to provide a more complete picture of an operator's state and whether or not they are likely to experience an "overload" condition. As a result, if a system can be developed that accurately assesses an operator's state, it may have the ability to contribute to the saving of lives in future aircraft systems. One way this ability may be realized is by notifying the operator that he is approaching a dangerous cognitive state in time for corrections to be made to reduce the workload level. Alternatively, onboard systems may be able to automate some functions that are normally controlled by the operator after an "overload" condition has been determined which would lead to a serious degradation of mission performance.

Current efforts at the Air Force Research Laboratory Flight Psychophysiology Laboratory (AFRL/FPL) at Wright-Patterson AFB are directed toward understanding the effects of mental workload through the use of brain electrical activity, heart rate, eye movements, and respiration patterns. These psychophysiological features are analyzed in attempt to identify changes in the mental workload of an operator in the laboratory, simulators, and aircraft. Previous research at AFRL/FPL has included studies of mental workload using simulated air traffic control [8,38,39], C-130 crewmembers during flight, and F-4 crewmembers during flight [50,52], among other studies. Current efforts include the study of mental workload of civilian pilots at the Wright-Patterson Aero Club and has included support from the Air Force Institute of Technology (AFIT) by Greene et. al. [20,21]. In addition, the AFRL/FPL is sponsoring this research effort by supplying data from a multi-task experiment, while the Air Force Office of Scientific Research (AFOSR) is sponsoring this effort via financial grants to AFIT.

In addition, similar mental workload research with operators in multi-task environments has been performed and is currently supported by other U.S. organizations such as the U.S. Army and NASA, along with much support in the European community. Specific examples of these efforts include studies performed by Caldwell et al. (U.S. Army) [10], Gevins et al. (NASA) [16,17], Galley [15], Jorna [27], Quartz et. al. [33], Roscoe [35], and Sirevaag et. al. [40].

To classify mental workload, traditional statistical techniques including ANOVA and jackknife methods have been used and are currently being used by many of the researchers [10,15,16,17,27,35,40,50,51,52]. Additionally, the emerging field of ANNs has been shown effective for pattern recognition and discrimination of various data sets [4,6,7,32,46,48]. ANNs are inspired by biological cognitive systems and have the ability to "learn" by adjusting the

weights of input connections to a given node, as will be described in Chapter 2. In short, ANNs have been found highly effective at classifying psychophysiological data including the numerous features resulting from EEG signals. Relevant work using ANNs to classify mental workload at AFIT and AFRL/FPL has been performed by Greene [19], Greene et. al. [20,21,23], and Russell et. al. [38,39].

1.3 Research Objectives

As stated, AFRL/FPL performs experiments in which psychophysiological data including heart-rate, eye-blinks, respiration, and electrical brain activity is collected. In addition to improving a classification model's output, identifying an optimal set of features has other advantages. First, current modeling efforts may include over 300 input features. While analysis of these features is possible in the laboratory, real-time analysis performed by an onboard system may not facilitate the necessary collection, processing, screening, and analysis of a data set of this magnitude. Additionally, if collection of physiological data in the cockpit is to be the future norm, universal data collection hardware that is minimally intrusive, yet reliable enough to obtain a robust set of features must be developed. At the present time, no known robust set of salient features has been identified to determine mental workload for use by different subjects. Thus, identification of a single robust set of salient features applicable for all individuals is desired.

To answer the proposed research question *"can one net fit all?"* two subsequent questions can be asked. First, do person-to-person psychophysiological variations necessitate an individual set of features for each person? Second, do person-to-person psychophysiological variations necessitate unique weighting of a common set of parsimonious salient features? The first question will be answered by determining if a single parsimonious set of features can be

identified, which yields a desired level of classification accuracy using various ANN models. The second question can be answered by comparing the classification results of separate ANN models, with models specifically trained for each test subject and a single model trained for a group of test subjects as a whole.

1.4 Research Methodology

As stated, physiological data was provided by AFRL/FPL from subjects who underwent testing in a multi-task environment. To collect this data, the Multi-Attribute Task Battery (MAT-B) was utilized in a controlled laboratory setting. MAT-B is user-interactive software developed by NASA for the research of human operator workload, and incorporates tasks analogous to activities that aircraft crewmembers perform in flight [11]. The specifics of MAT-B and the simulated workloads will be explained in greater detail in Chapter 3.

Several saliency metrics are available for use to determine an optimal, parsimonious set of features for use by an ANN. Among these are Ruck's Saliency metric [36] and Tarr's Saliency metric [46], which are presented in Chapter 2. Additionally, a relatively new saliency measure based on the saliency of an input feature to that of injected noise is presented in Chapter 2. This Signal-to-Noise (SNR) saliency measure was first demonstrated by Sumrell [45] and has been utilized in similar workload classification efforts by Greene [19,20,23]. In addition, the SNR saliency measure has the advantage of use "on-the-fly" while training an ANN, and will be utilized by this research.

While the specific methodologies of this research are included in Chapter 4, a quick overview of the approach is as follows:

- Use SNR saliency screening for each individual test subjects to determine a parsimonious set of salient features for each individual. Calculate classification accuracy of each subject using multiple runs of ANN models with each individual's parsimonious set of salient features.

- Use SNR saliency screening for all test subjects as a group to determine the parsimonious set of salient features for the group. Calculate classification accuracy using multiple runs of one ANN with the parsimonious set of salient features.

- Using the same validation sets of data compare classification accuracy as a measure of effectiveness to determine if the group ANN is significantly different than each individual's trained ANN.

- Compare classification accuracy and the parsimonious sets of salient features to those obtained using a linear multivariate discriminant approach.

1.5 Scope of Research

As stated, the primary goal of this research is to identify if a robust parsimonious set of physiological features for workload classification exists, with a secondary goal of determining if *"one net can fit all."* Additionally, this research effort provides the following:

- Further development of SNR saliency screening for ANNs

- Investigation of changes to the current psychophysiological data preprocessing which may lead to better classification accuracy.

- Creation of a CD archive of all data provided by AFRL/FPL

- Creation of a CD archive of all processed psychophysiological data

- Development of *Matlab* m-files to perform discriminant analysis

- Development of *Matlab* m-files to perform ANN saliency screening

- Modification of *Matlab Neural Network Toolbox* m-files to support the specific ANN modeling efforts of this research

II. Background and Literature Review

This chapter provides a review of the literature concerning four primary areas in this research effort. First, a description of feedforward multilayer perceptron (MLP) artificial neural networks (ANNs) and relevant definitions is presented as a foundation for the primary models to be used in this research. Next, a review of ANN salient feature selection is presented. Following the review of ANNs and saliency screening, multivariate discriminant analysis is presented as a means to compare ANN models and ANN input saliency screening. Finally, psychophysiological features are reviewed as they will be the variables used by both linear discriminant and nonlinear ANN models to assess mental workload in a multi-attribute task environment.

2.1 Feedforward MLP ANNs

This literature review starts with a brief overview and history of ANNs, defines relevant terms to establish consistency throughout this thesis, and proceeds to a complete description of a perceptron and a feedforward MLP ANN. The architecture of feedforward MLP ANNs is then addressed, the backpropagation training algorithm is described, and finally other relevant issues such as typical data transformations and variations to the backpropagation algorithm are discused.

2.1.1 Overview. ANNs are inspired by biological cognitive systems and have the ability to "learn." Learning is accomplished by providing feedback under supervised training to adjust model parameters to provide more accurate output. ANNs also utilize

parallel computing similar to biological neural systems in which information is stored and calculations are performed by an architecture with parallel structure [7,32].

The first perceptron model was first introduced in the late 1950s by Frank Rosenblatt, and is considered to be a two-layer feedforward ANN [7]. The input layer is not counted, the first layer contains fixed threshold logic functions, and the second layer provides the network output and has connecting weights that are trainable, as can be seen in Figure 2-1.

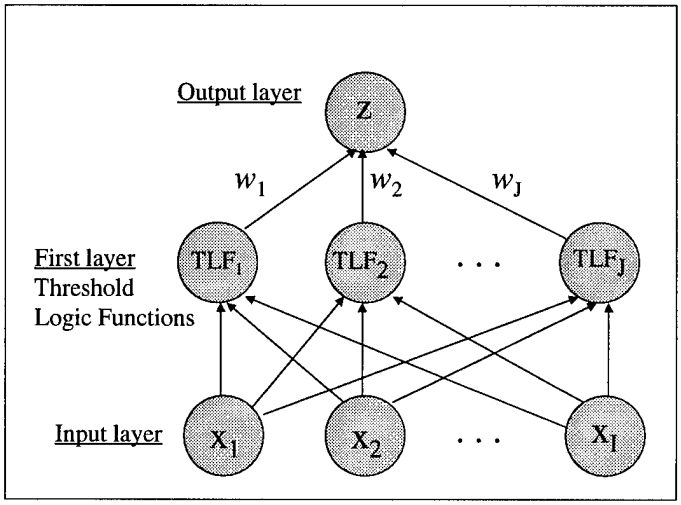

Figure 2-1. Rosenblatt's Perceptron.

With initial heavy criticism of the perceptron for its limitations, research and development in the field of ANNs progressed slowly for the next few decades. By the late 1980s new developments and the application of theory led to a new optimism for potential commercial and biological modeling research uses of ANNs. The Department of Defense (DoD) also sponsored a substantial study of ANNs under the auspices of MIT's Lincoln Laboratory from 1987 to 1988 with academic, industry, and government participants [12]. The goals of this study were to determine the current neural network

technology base, to identify technology requirements, and to identify potential applications for neural networks in DoD systems, all of which were achieved. Additionally, the increasing growth rate and support of the field can be seen in the increased number of conferences and refereed journals. Finally, as commercial applications continue to mature, funding from private industry will continue to provide the necessary support of the field. Thus, the field of ANNs has only recently developed, with the possibility of numerous applications including complex pattern recognition. One specific application where ANNs have been found highly effective, is at recognizing patterns in psychophysiological data. This classification of patterns includes analyzing numerous channels of electroencephalogram (EEG) data which records electrical activity of the brain at multiple locations on a person's head.

2.1.2 Definitions.

- **Artificial Neural Network (ANN).** An information processing system that operates on inputs to extract information and produces outputs corresponding to the extracted information [12].
- **Architecture.** The topological arrangement of neurons, layers, and connections, which defines the set of modeling equations available to the ANN [53].
- **Backpropagation.** A learning algorithm for updating weights in a feedforward MLP ANN that minimizes the mean squared mapping error [12].
- **Epoch.** A complete presentation of the data set being used to train the MLP, or equivalently called a training cycle [4].

- **Feature.** In neural networks, features refer to the input vectors of information which are presumed to have some relation that may be helpful in distinguishing the various output classes [4].
- **Feedforward.** Multilayer ANNs whose connections exclusively feed inputs from lower to higher levels. In contrast to a feedback or recurrent ANN, a feedforward ANN operates only until all the inputs propagate to the output layer. An example of a feedforward ANN is the MLP [12].
- **Hidden Units.** The processing elements in MLP ANN that are not included in the input or output layers. This is the part of the neural network located between the input and output where complex problem solving occurs [12].
- **Learning Algorithm.** The equations used to modify the weights of processing elements in response to input and output values [12].
- **Neuron.** The fundamental building block of an ANN. Normally, each neuron takes a weighted sum of its inputs to determine its net input. The net input is then processed through its transfer function to produce a single-valued output which is broadcast to 'downstream' neurons [53].
- **Single-layer Perceptron.** A type of ANN algorithm used in pattern classification problems that is trained using supervision. Connection weights and thresholds can be fixed or adapted using a number of different algorithms [12].
- **Supervised Training.** A method of training adaptive ANNs that requires a labeled training data set and an external teacher. The teacher knows what the desired response is and thus can provide responses for correct or incorrect classification by the network [12].

- **Weight.** A processing element (or neuron or unit) need not treat all inputs uniformly. Processing elements receive inputs by means of interconnects (also called 'connections' or 'links'); each of these connections has an associated weight which signifies its strength. The weights are combined to calculate the activations [12].

2.1.3 Description of a Feedforward MLP ANN. Within a MLP ANN, a perceptron receives a weighted sum of M features and a bias term. The perceptron then performs a mathematical transformation on this weighted sum. The transformation then serves as the perceptron's output. Figure 2-2 is an example of a single perceptron with a bias term included. From the figure, data is fed upward into the perceptron through input nodes x_1 through x_M, with associated weights w_i for each input. The perceptron then proceeds to sum across the weighted inputs, adding the bias term, and produces an activation output value which is identified below.

$$\text{Output} = f\left[\left(\sum_{I=1}^{M} w_i x_i\right) + \theta\right] \tag{2-1}$$

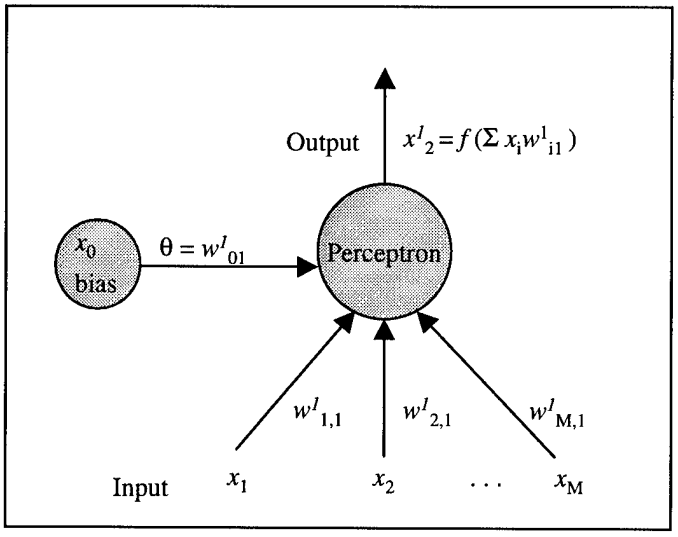

Figure 2-2. Single Perceptron with Bias.

The bias or threshold, is an additional node added to each layer of a MLP ANN, with a constant input value of one. In other words, the bias provides a constant input to each hidden and output node regardless of the input features. This constant input is equal to the bias layer weight connecting into each node.

Desirable transformations functions typically squash all input values into a predetermined range (normally –1 to 1). Examples of such functions are provided in the following figures and include a hard limiter, threshold logic as used by Rosenblatt's first perceptron, hyperbolic tangent, and a sigmoid function.

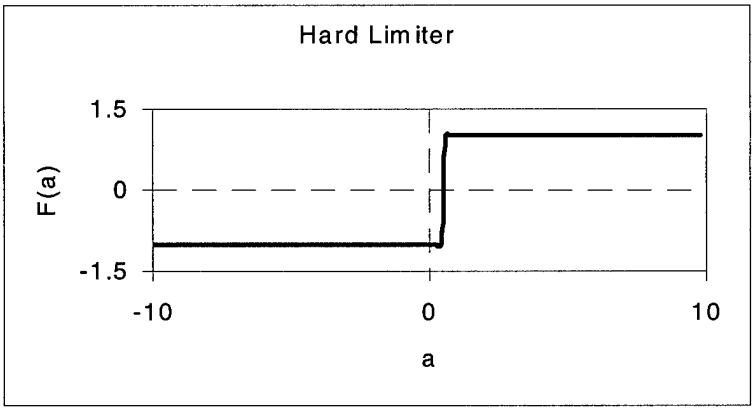

Figure 2-3. Hard Limiter Function.

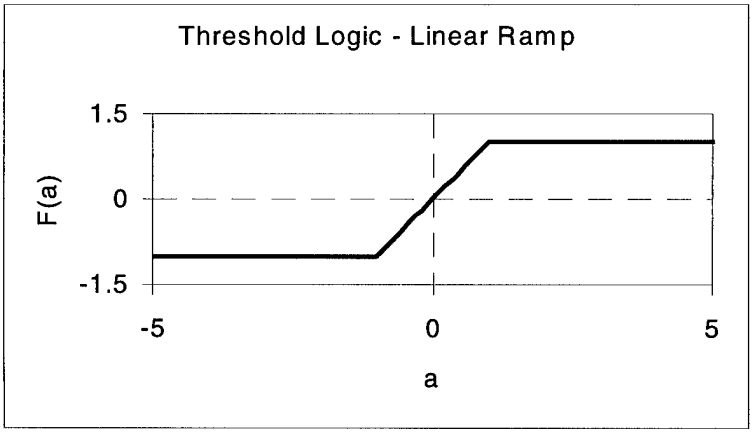

Figure 2-4. Threshold Logic Function.

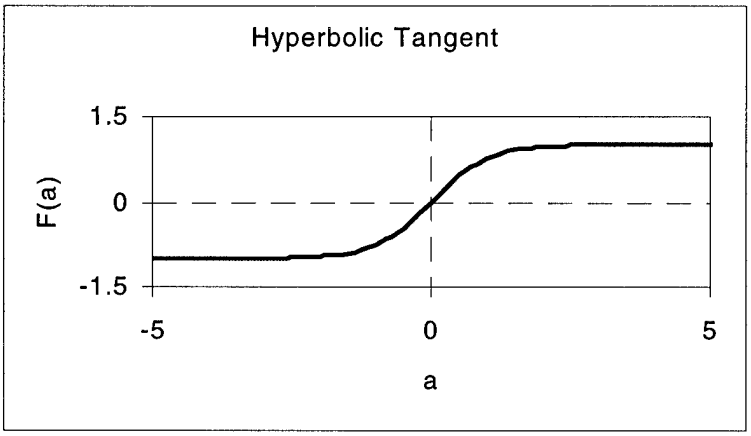

Figure 2-5. Hyperbolic Tangent Function.

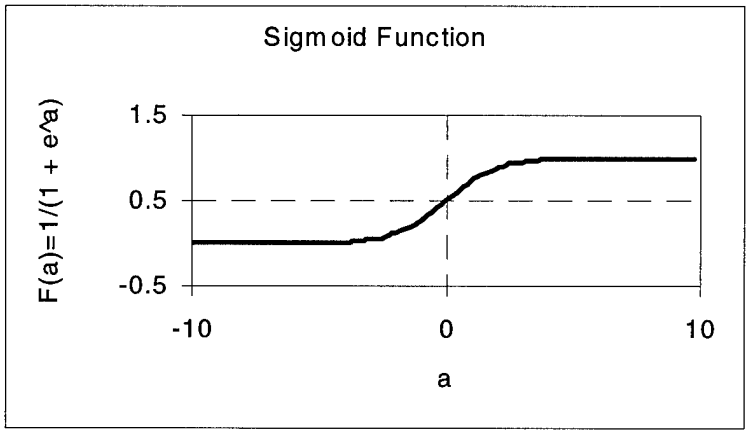

Figure 2-6. Sigmoid Function.

While the hard limiter and threshold logic are linear in nature, the hyperbolic tangent and sigmoid functions are non-linear and provide for a continually differentiable function that is more desirable. Figure 2-7 below represents a fully connected MLP ANN.

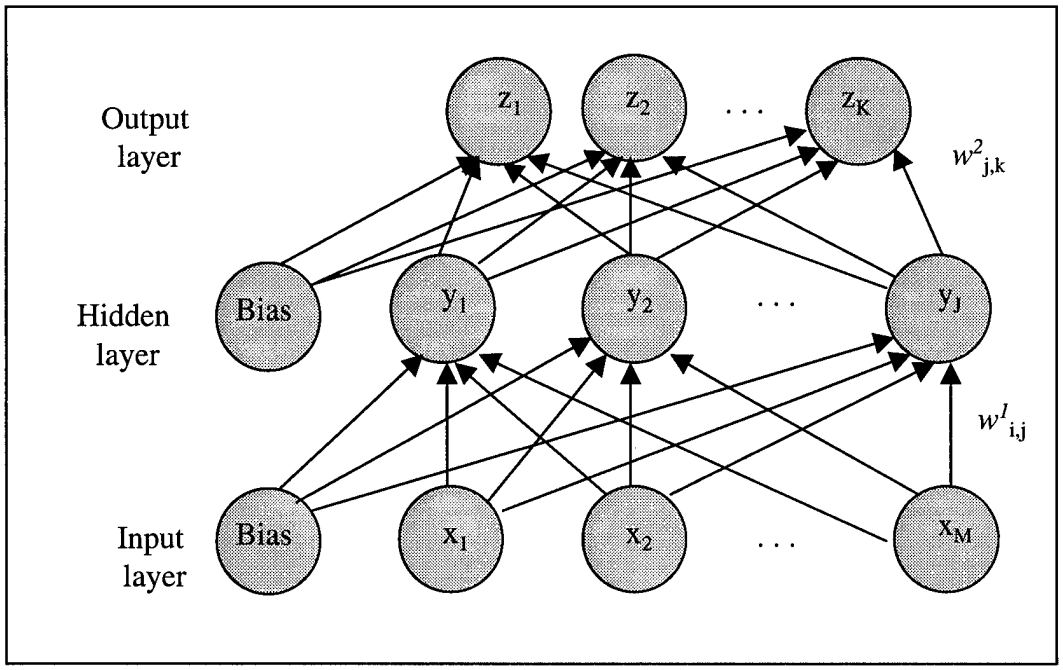

Figure 2-7. MLP ANN with Bias.

The output from such a MLP ANN for the nth input vector (z^n) can be computed as follows:

$$k\text{th neural network output} = z^n_k = f\left(\sum_{j=1}^{J} w^2_{j,k} x^1_j\right), \text{ where} \qquad (2\text{-}2)$$

- J is the number of hidden nodes.
- $f(a) = 1/(1 + e^{-a})$ for sigmoidal activation functions.
- $f(a) = a$ for linear activation functions.
- $w^2_{j,k}$ is the weight from hidden node j to output node k.
- x^1_0 is the hidden layer bias term and is set equal to 1.
- $x^1_j = f(\Sigma w^1_{ij} x^n_i)$ is the output of hidden node j and is summed from i=1 to M.
- M is the number of input features.
- w^1_{ij} is the weight from input node i to hidden node j.
- x^n_0 is the input layer bias term and is set equal to 1.
- x^n_i is the ith input feature of the nth input vector.

2.1.4 Architecture. Selecting an appropriate architecture is important to create an efficient ANN. Choosing too few hidden nodes may result in a network that does not converge in a reasonable period of time or possibly not at all. Alternatively, if too many (redundant) nodes are chosen, a network's ability to characterize new data may be reduced [34]. For any given ANN application, multiple architectures can normally be utilized which will perform similarly and will provide equivalent classification levels. An important issue in architecture selection then becomes choosing the most efficient network that will provide a desired level of classification. To minimize computations and training time, an ideal network will have the minimum number of nodes required to achieve a specified classification accuracy [34]. The following four variables are used to define the architecture of a MLP ANN.

- **Input Nodes.** The input layer contains one node for every feature.
- **Output Nodes.** The output layer normally has one node for every class the model is attempting to identify.
- **Hidden Layers.** Although any number of hidden layers can be used, one hidden layer of perceptrons will be used in this research effort. Cybenko and Hornik have shown that one layer of hidden nodes (two layers of weights) is sufficient for any multivariate approximation problem [42,46]. A desired accuracy can then be obtained by choosing a sufficient number of hidden nodes and utilizing a sigmoidal or appropriately smooth alternative activation function.
- **Hidden Nodes.** With the number of hidden layers set to one, the primary variability between the architecture of a multilayer perceptron with predetermined input and output features then becomes the number of hidden nodes. To date, a deterministic

algorithm has not been developed for selection of the most efficient number of hidden layers and nodes. Hidden layer architecture selection is typically heuristically chosen and is more of an art form. As Ruck states, "Rigorous mathematical techniques have not been developed to determine the appropriate number of hidden layers or the number of nodes in those layers for a given problem"[36]. Steppe, adds that, "Although the number of required hidden nodes is unknown in advance, a reasonable number of hidden nodes is often determined by a trial and error process or by more sophisticated methods" [42]. Some of these more sophisticated methods utilize saliency screening and include algorithms developed by Steppe and Bauer [43]. Other algorithms used include evolutionary programming as proposed by Sakar and Yegnanarayana which minimizes a mean square error function, and an iterative construction algorithm by Rathburn and others which adds hidden nodes sequentially until all data pairs are separated. Additionally, Rizzo has proposed two methodologies to select an appropriate number of nodes. These methods include utilizing the Akaike Information Criterion (AIC) and the use of the Signal-to-Noise saliency measure. While both methods reported the ability to recommend reductions in the number of hidden nodes, resulting in more efficient networks, the AIC selection methodology appeared conservative in nature, while the SNR methodology remained very heuristic in nature [34].

Additional considerations for an MLP ANN include the transformation of raw feature data, the learning rate step-size to be used, whether or not to use momentum and the appropriate momentum rate, what ranges of weight initialization to use, and the number of epochs to train through. Many of these considerations will be addressed in the following discussion of backpropagation.

2.1.5 Backpropagation. Training algorithms are the rules by which perceptrons can update their weights as data is presented. Backpropagation is probably the most popular algorithm for finding an MLP ANN's weight parameters. In an MLP ANN with backpropagation, supervised learning is used to approximate a unidirectional mapping from an m-dimensional input space $\mathbf{R}^{\mathbf{M}}$ (where M is the number of input features) to an k-dimensional output space $\mathbf{R}^{\mathbf{K}}$ (where K is the number of output classes) [53]. In doing so, backpropagation is an iterative gradient descent algorithm requiring sample input data with correct mappings to the output space.

Two methods of updating weights can be utilized. The first method is an instantaneous update that examines the gradient of the error surface after the network incorporates each training vector. The second method utilizes a batch update that only examines the gradient after the network has seen all the training vectors [36]. Presented next is an algorithm for calculating the instantaneous back propagation as presented by Steppe [42]:

1. Randomly partition data into a training, training-test, and validation sets.
2. Normalize the feature input data.
3. Initialize the weights to small random values.
4. Present the network with a randomly selected vector from training set, denoted $\mathbf{x}^{\mathrm{p}}$.
5. Calculate the network output $\mathbf{z}^{\mathrm{p}}$ associated with the *p*th training vector.
6. Update the weights.
7. If test set error rate does not indicate sufficient convergence return to step 4.

In step 1, the feature data is randomly divided into two sets. The first set is the training set that is used to calculate training weights while simultaneously evaluating network performance on a test set. The validation set is used as an independent source of data for assessing model adequacy at predicting future responses. This use of the three sets is desired as neural nets have the capability to "memorize" noise in the data. The use

of noisy training data can then lead to problems similar to over-fitting with regression models containing high order terms [49]. Thus, a network's performance may be very good for the training set, while very poor for the training-test or validation set. A single data set can be divided into adequate training and validation sets using an approximate 3:1 ratio. The test set can then be divided between training and training-test sets by utilizing a 2:1 ratio. The resulting partition of training : training-test : validation is then approximately 2:1:1. Alternatively other ratios may also be used such as a 40/30/30 percentage split of the data or simply assigning an equal number to all three sets. Ultimately, each application's goals and the number of exemplars available may be the deciding factors in how the sets are formed.

Data can be input into a neural network as raw data or after some appropriate transformation has been made. In step 2, data is normalized so that all features have the same range (usually –1.0 to 1.0 or 0.0 to 1.0). Or, statistical standardization can be performed that normalizes each feature to a mean of 0.0 and a variance of 1.0 [36,42]. In order to keep the validation and test sets as independent as possible, Steppe suggests normalizing the two sets independently in order to keep all normalization information separate [42].

In step 3, weights are initialized for all connections. These weights are chosen randomly and are normally within the range of –1.0 to 1.0. For many applications, a range of –0.5 to 0.5 is utilized [42], although when performing SNR feature screening initial weights closer to 0.0 may be more desirable [19,21].

In step 4, a randomly selected feature vector $\boldsymbol{x}^p$ is input into the current neural network. With p indicating that $\boldsymbol{x}$ is the pth vector of the training set. Step 5 then is used

to calculate the output from the network using the summations of sigmoid functions and the current weights as defined in equation 2-2.

In step 6, the instantaneous output error $\varepsilon^p{}_o$ associated with $\boldsymbol{x}^p$ is calculated using the pth vector of outputs denoted as $\boldsymbol{z}^p{}_k$ and the corresponding vector of desired outputs $\boldsymbol{d}^n{}_k$. Where p represents the pth input exemplar vector of data, and k represents the number of output nodes, which is usually equal to the number of classes. Instantaneous network output error $\varepsilon^p{}_o$ is the squared error associated with the pth exemplar and is given as:

$$\varepsilon^p{}_o = \sum_{k=1}^{K} (d^p{}_k - z^p{}_k)^2 \tag{2-3}$$

where K is the number of output nodes, $d^p{}_k$ is the desired output associated from the pth exemplar at the kth output, and $z^p{}_k$ is the observed network output produced from the pth exemplar at the kth output. The gradient decent step direction is then determined by taking the partial derivative of $\varepsilon^p{}_o$ with respect to the weight parameters. Thus, weights are updated as follows:

$$\text{Hidden layer to output layer weights } (w^2{}_{jk})^{new} = (w^2{}_{jk})^{old} + \eta\delta^2{}_k x^1{}_j \tag{2-4}$$

$$\text{Input layer to hidden layer weights } (w^1{}_{ij})^{new} = (w^2{}_{ij})^{old} + \eta\delta^1{}_j x^n{}_i \tag{2-5}$$

Where,

- $(w^2{}_{jk})^{new}$ is the updated weight from hidden node j to output node k.
- $(w^2{}_{jk})^{old}$ is the old weight from hidden node j to output node k.
- $(w^1{}_{ij})^{new}$ is the updated weight from input node i to hidden node k.
- $(w^1{}_{ij})^{old}$ is the old weight from input node i to hidden node k.
- η is the training step size.
- $\delta^2{}_k = (d^n{}_k - z^n{}_k) z^n{}_k (1 - z^n{}_k)$ if $f[\cdot]$ is a sigmoid function.
- $\delta^2{}_k = (d^n{}_k - z^n{}_k)$ if $f[\cdot]$ is a linear function.
- $\delta^1{}_j = x^1{}_j (1 - x^1{}_j) \Sigma\, \delta^2{}_k (w^2{}_{jk})^{old}$ for k=1 to K, if $f[\cdot]$ is a sigmoid function.

- $\delta^1_j = \Sigma\, \delta^2_k (w^2_{jk})^{old}$ for k=1 to K, if $f[\cdot]$ is a linear function.
- d^p_k is the kth desired output for the pth exemplar.

The step size, η, can be constant or variable. The larger the rate is, the bigger each step in the gradient search will be. But, if the learning rate is made too large, the algorithm will become unstable, and if the learning rate is set too small, the learning algorithm will take too long to converge [13]. Additionally, Steppe states from White, that "a constant learning rate is inefficient because the random influences in the input will result in random fluctuations in the weight vector preventing backpropagation from ever settling down to the optimal weight vector" [42]. More efficient, proposed learning rate functions are seen to decline over time. Proposed functions for η include learning rate functions that are inversely proportional to the number of epochs or the log of the number of epochs.

Backpropagation can also be implemented using a gradient descent that includes momentum. Momentum allows a network to respond not only to the local gradient, but also to recent trends in the error surface. Much like a low pass filter, momentum allows the network to ignore small features in the error surface. Without momentum an MLP ANN may get stuck in a shallow local minimum, while with momentum an MLP ANN can slide through such a minimum [13]. Momentum can be added to backpropagation learning by making weight changes equal to the sum of a fraction of the last weight change and the new change suggested by the backpropagation rule. The modified backpropagation equations are then as follows:

Hidden to output layer weight:

$$[w(t+1)^2_{jk}]^{new} = [w(t)^2_{jk}]^{old} + \eta\delta^2_k x^1_j + \beta\Delta[w(t-1)^2_{jk}]^{old,\ old} \qquad (2\text{-}6)$$

Input to hidden layer weight:

$$[w(t+1)^1_{ij}]^{new} = [w(t)^2_{ij}]^{old} + \eta\delta^1_j x^n_i + \beta\Delta[w(t-1)^2_{ij}]^{old,\,old} \qquad (2\text{-}7)$$

where,

- β is the momentum constant that determines the effect of past weight changes on the current weight change
- $[w(t+1)]^{new}$ is a weight at epoch t+1
- $[w(t)]^{old}$ is a weight at epoch t
- $\Delta[w(t-1)]^{old,\,old} = [w(t)^{old} - w(t-1)^{old,\,old}]$, weight change from epoch t –1 to epoch t
- t is the training epoch

Thus, with a momentum constant rate β of zero, a weight change is based solely on the gradient as before. Alternatively, a momentum constant of one will result in a new weight change that is set equal to the last weight change plus the current gradient. As a final note, the momentum should never exceed one as this implies an exponential impact on training [49]. In summary, the learning rate step size η determines the magnitude of the next step to take while implementing a gradient descent backpropagation learning algorithm. While the momentum constant β determines how much the direction will change from the previous step taken [45].

2.2 Saliency Measures and Screening for ANN Feature Selection

With multiple measures of heart-rate, respiration, and eye-blink, and the potential for over 300 features of EEG data collected at 60 locations, reduction in the number of features selected to train an ANN is highly desired. By reducing the number of features, tasks involved in data collection, management, and analysis can be accomplished much more efficiently. Additionally, some of the many input features may contain excessively large amounts of noise or may not have any significant relationship to mental workload. In these cases, any neural network model used has little chance of accurately classifying

the data. Thus, to avoid such cases of *"garbage-in, garbage-out"* only salient features are desired as input data. Review of current literature has identified the application of three saliency measures to aid in the determination of a parsimonious set of salient features. Specifically, three saliency measures have been utilized in recent efforts to reduce psychophysiological features [19,21,22,38,39] and are as follows:

- Ruck's saliency measure
- Tarr's saliency measure
- Signal-to-Noise Ratio (SNR) saliency measure

These saliency measures can be implemented using different algorithms. Relevant work in this area includes screening methodologies developed by Belue and Bauer [5], Steepe and Bauer [43], along with a SNR screening application as first demonstrated by Sumrell [45] and latter applied to real world problems by Greene [19].

2.2.1 Ruck's Saliency Measure. Ruck's saliency measure is built from the partial derivatives of network outputs, $\mathbf{z}_k$, with respect to feature inputs, $\mathbf{x}_m$, using a trained network. This measure uses the sum of the partial derivatives of the network outputs with respect to the entire M-dimensional input space, $\mathbf{R}^M$, where m = 1, 2,...,M. Numerically, Ruck's saliency measure can be calculated using pseudo-sampling. For feature i, this is computed as follows:

$$\hat{\Lambda}_i = (1/K)(1/R)\sum_{k=1}^{K}\sum_{r=1}^{R} \left| \partial z_{k,r}/\partial x'_{i,r}(\mathbf{x}'_r, \mathbf{W}) \right| \tag{2-8}$$

Where the M-dimensional input space is now divided into r = 1, 2,...,R uniformly spaced bins, and K is the number of output classes, and **W** is the weights of a trained ANN. The partial derivatives are taken at the midpoints of each range bin denoted as $\mathbf{x}'_r$ for r =

1,2,...,R. Since the partial derivative of the outputs is dependent on **W** with the final weights affected by initial values, Ruck's saliency measure is typically calculated over a number of independently trained networks [36]. For implementation of the Bauer-Belue or Steppe-Bauer screening methodologies, average values of the saliency measure from 10 to 30 separately trained ANNs is suggested.

2.2.2 Tarr's Saliency Measure. Tarr's saliency measure does not use partial derivatives, but uses various norms of the weights between input and hidden nodes. The equation for Tarr's metric of feature i is as follows [46]:

$$\tau_i = \sum_{j=1}^{J} (w_{i,j}^{1})^2 \tag{2-9}$$

where τ_i is the Tarr saliency metric for feature i, J is the number of hidden nodes, $w_{i,j}^{1}$ is the first layer weight between input node i and hidden node j. This equation for Tarr's saliency as defined above is simply the sum of the squared weights between input node i and all hidden nodes 1 through J.

2.2.3 Signal-to-Noise Ratio (SNR) Saliency Measure. SNR is a saliency measure that is similar to Tarr's metric in that both rely on the sum of squared first layer weights [45]. The SNR metric is different from both Ruck's and Tarr's metrics because it directly compares the saliency of a feature to the saliency of an injected noise feature. The SNR saliency metric is computed as follows:

$$SNR_i = 10 \cdot \log_{base10}\left(\frac{\sum_{j=1}^{J} \left(w_{i,j}^{1}\right)^2}{\sum_{j=1}^{J} \left(w_{N,j}^{1}\right)^2} \right) \tag{2-10}$$

where SNR_i is the value of the saliency metric for feature i, J is the number of hidden nodes, $w^1_{i,j}$ is the weight from node i to node j, and $w_{N,j}$ is the first layer weight from the noise node N to node j. The injected noise is created as a Uniform(0,1) distribution. The scaled logarithmic transformation of the ratio converts the saliency measure to a decibel scale. Like Tarr's and Ruck's saliency metrics, the SNR metric can be used to rank order input features. Additionally, if a given feature is not relevant to a neural network's output, the updates of the first layer weights from that feature's node should be random and fluctuate close to zero. On the other hand, if a given feature is relevant, the weights should be adjusted in a constant direction until error in the network is minimized. Thus, the resulting SNR saliency metric should be significantly larger than 0.0 for salient features and close to 0.0 for non-salient features. The greatest potential of the SNR saliency metric results from its comparison of the saliency of each feature to that of a baseline noise feature. This allows for the SNR metric to be calculated and used at any time during network training which can not be done with the Ruck or Tarr saliency measure.

A proposed methodology to implement the SNR saliency screening in a multilayer perceptron is outlined by Sumrell [45] as follows:

1. Add a noise feature, x_N, to the original set of features.
2. Begin training of the neural network.
3. Interrupt training after the salieny metric values have stabilized.
4. Identify the feature with the lowest SNR value and remove it from further training.
5. Continue training the neural network.
6. Repeat steps 3, 4, and 5 until all of the features in the original set have been removed.
7. Compare the reaction of the test set classification error rate to the removal of the individual features.
8. Retain the first feature whose removal caused a significant increase in the test set classification error rate, as well as all features that were removed after that first salient feature.

Because the SNR saliency measure relies heavily upon the feature ranks being consistent from one run to the next, Sumrell investigated the robustness of the SNR saliency metric over a designed region of neural net architecture. Factors in the design space included the number of hidden nodes, the learning rate step size, and the momentum rate used for training. Using two separate data sets and networks, the SNR saliency metric was found to be robust to changes in both the number of hidden nodes and the learning rate step size. On the other hand, the momentum rate did appear to influence the ability of the SNR metric to correctly rank order features. Thus, for use of the SNR metric, Sumrell's investigation recommends neural network architectures with N to 3N hidden nodes (where N is the number of input features), a learning rate step size between 0.1 and 0.9, and a momentum rate between 0.1 and 0.5.

2.3 Linear Multivariate Classification

As mentioned, linear models will be developed and used as a benchmark for assessing the salient features selected and the classification accuracy obtained when using ANN models. Specifically multivariate discriminant analysis will be performed as a means to classify data into exclusive workload levels. The following sections include a description of multivariate discriminant methodology, feature selection, and a means for comparing the models.

2.3.1 Multivariate Discriminant Analysis. Discriminant analysis as defined by Dillon and Goldstein is "...a statistical technique for classifying individuals or variables into mutually exclusive groups on the basis of a set of independent variables" [14]. Specifically, the goal of multivariate discriminant analysis is to derive linear

combinations of independent input variables that will distinguish observations between *a priori* defined classification groups in a manner that minimizes misclassification error rates.

In particular, one discriminant analysis algorithm will be applied in this research that uses psychophysiological features to classify mental workload. The method utilized will determine a set of quadratic discriminant scores, d^Q, for each input observation **x**. A separate d^Q score is determined for each of k classification populations. Each score is derived from the probability of the observation belonging to a specified multivariate normal distribution. The observation is then assigned to the population with the largest associated score, which corresponds to the greatest multivariate normal distribution probability. The d_k^Q score associated with a given observation and population can be computed as follows:

$$d_k^Q(\mathbf{x}) = -\frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\mathbf{x}-\mu_k)'\Sigma_k^{-1}(\mathbf{x}-\mu_k) + \ln P_k \qquad (2\text{-}11)$$

where Σ_k is the estimated covariance matrix for population k, μ_k is the sample mean for population k, and P_k is the *a priori* probability the observation is from population k.

2.3.2 Feature Selection for Multivariate Discriminant Models. Like ANN models, the efficiency and effective classification accuracy of discriminant models can be improved by using a parsimonious set of salient input features. Additionally, features found salient for use in a linear model are likely to be desired for use in a non-linear model. One argument supporting the desired use of terms that are linearly salient is the sparsity of effects principle, where low order terms tend to dominate, as can be seen in a Taylor series expansion [9]. To screen input features, univariate F-values and

standardized discriminant function coefficients can be used. Yet, both measures may provide an inaccurate picture of a variable's importance because they ignore the interrelationships between the variables. Resultant problems include the confounding of these quantitative measures by variables that contribute redundantly to the overall classification [14]. In contrast, a discriminant loading gives the simple correlation of each input variable with the discriminant function and does not suffer from potential problems associated with multicoliniarity of input variables. As stated by Dillon, discriminant loadings, "...are less subject to instability caused by predictor intercorrelations, and thus tend to be more useful in an interpretation than standardized discriminant weights." While Dillon and Goldstein support the use of discriminant loadings to facilitate input feature interpretation, they do not provide the methodology to compute consistent discriminant loadings. Fortunately, Bauer [3] does provide a single equation to compute the discriminant loadings, CORR($\mathbf{X}$,$\mathbf{b}'\mathbf{X}$), which is as follows:

$$\text{CORR}(\mathbf{X},\mathbf{b}'\mathbf{X}) = (\mathbf{b}'\text{COV}(\mathbf{X},\mathbf{X})\mathbf{b})^{-1/2}\,\text{CORR}(\mathbf{X},\mathbf{X})D_X^{-1/2}\mathbf{b} \qquad (2\text{-}12)$$

Where $\mathbf{X}$ is the matrix of input observations, $\mathbf{b}'\mathbf{X}$ defines a linear boundary that can be obtained from equation 2-11, $\mathbf{b}$ is the vector of discriminant weights, and D_X is the matrix of diagonal elements of COV($\mathbf{X}$,$\mathbf{X}$).

2.3.3 Comparision of Classification Models and Feature Selection. The prior discussion in this chapter has concentrated on the background to determine the parameters for both a linear and a nonlinear classification model, along with feature selection for both. While optimizing the set of input features and model parameters is important, also important is ensuring an accurate measure of a model's performance can

be determined. By determining an unbiased, accurate measure of performance, the relative value of a model can be assessed. Additionally, each model's performance may now be compared fairly. To assess a classifier's performance, the percentage of correct classifications or the percentage of incorrect classifications (error rate) can be used as a standard measure of effectiveness. Following is a review of three methods to compute the error rate of a model and include resubstitution, data splitting, and leave-one-out methods.

The *resubstitution* method utilizes the same set of data to design a classification model and to assess its accuracy. This method produces an *apparent* error rate that typically underestimates the expected or *actual* error rate obtained when new data is classified by the model. As stated by Dillon, "The estimates are consistent, but can be severely optimistically biased." [14]

In contrast to resubstitution, the *data splitting* procedure requires dividing the original data set into two separate subgroups. One subgroup is used to train a model, while the second group is used as an independent validation set to determine the expected error rate [3]. Criticism of this methodology includes its rather inefficient use of data in which information contained in the validation set is not utilized to determine the model parameters. Thus, large data sets are required to ensure models can be optimally determined and the error rate can be accurately assessed. In addition, variations of this method include splitting the data set into K random subgroups of test and validation, where a separate model and error rate can be determined K times.

Lastly, the *leave-one-out* or cross-validation method as originally proposed by Lachenbruch makes use of all available data without serious bias in estimating the error

rate [14]. With a data set of n observations, the error rate is determined by training a model using all but one observation. The left-out observation is then classified by the trained model. This procedure is repeated n times with each observation being left-out. The estimated error is then given by the percentage of misclassified left-out observations. Due to its computational intensity, this procedure may not be practical for use with complex models such as MLP ANNs, but it may be the preferred method if only a limited amount of data is available.

In summary, the data splitting and leave-one-out methodologies are superior to the resubstitution methodology as they provide an estimate of the expected *actual* error rate. Thus, to accurately assess and compare models, one of these two methods should be used. Additionally, the same validation data set should be used to obtain an estimate of error when comparing two or more different models.

2.4 Psychophysiological Features

In-depth literature review has identified the use of several psychophysiological variables utilized to classify mental workload in a multi-task environment. The study of mental workload in multi-task environments is clearly a concern to U.S. Department of Defense agencies and to NASA, as evidenced by their many funded studies in the area [1,10,11,15,16,17,18,19,21,20,23,29,38,39,50,52]. Additionally, the "Commission of European Communities expressed its interest in 'Human Performance in Transport Operations' by initializing a multinational programme in 1982" [25]. A major project in this effort was aimed at describing 'performance decrements,' with a resulting Special Issue of *Ergonometrics* in 1993 that is devoted to psychological measures in transport systems [25]. The transport systems studied include both automobiles [15] and aircraft

[27,35,40,50] in actual and simulated scenarios with studies performed in the US and in the European Community. Additionally, recent research by NASA, US Army, US Air Force, and other researchers [16,17,19,20,21,22,23,29,31,33,38,39] have used ANNs as a tool to classify psychophysiological data. As found in this literature review, six primary psychophysiological features can be used to assess mental workload and are as follows:

- Measures of Heart Rate
- Measures of Respiration
- Measures of Eye-Blink Activity
- Measures of Brain Activity
- Measures of Hormone Levels
- Measures of Electrodermal Activity

2.4.1 Measures of Heart Rate. Heart rate measures, alternatively referred to as electrocardiography (ECG) data, have been frequently used to assess workload in multi-task settings. The use of heart rate to measure pilot responses to flight was reported as early as 1932 [51] and is still used today. In general, increases in heart rate have been associated with increases in mental workload. Two of the more significant general findings are as follows. First, heart rate can be used to provide a measure of flight-segment workload including aspects such as take-off, landing, cruise, angle of descent to landing, etc. [18,27,35,50,51,52]. Second, increased heart rate provides a measure to discriminate responsibility of handling the aircraft between crewmembers and not just the stress of flying an aircraft [50, 52]. A second measure of heart rate is the variability of beat-to-beat heart rate. The main idea is that the extent of the normally found beat-to-beat variability decreases with increased mental workload, and provides additional information not provided by heart rate alone [51]. Two potential problems have been observed in the use of heart rate data to assess cognitive demands. First, heart rate

variability may decrease with age which would invalidate its use by many test subjects and for comparison between them. Second, simulated flight segments may not provide a difference in heart rates, while actual flight shows significant differences. Thus, a subject's heart rate response in a simulated environment may be different than in the same real world scenario.

2.4.2 Measures of Respiration. As may be expected, respiration rates tend to become more rapid as workload conditions, either physical or mental, increase. Although limited studies have used respiration rate as a measure of mental workload, those performed have reported increased respiration during periods of higher cognitive demand in both flight and air traffic control scenarios [8,50]. Additional studies demonstrating increased respiration rates in more demanding flight segments are referenced by Wilson and Eggemeier [51]. Overall, while an increase in respiration rate appears to be a good indication of increased workload conditions, collection and interpretation of respiration data can be difficult in real world scenarios. Because speech disrupts the pattern of breathing, applying measures of respiration to air traffic control or aircraft pilot scenarios where voice communications are normally required can prove to be difficult. Also of interest, is the suggested use of voice analysis as a measure of operator workload, where both fatigue and increased workload are thought to cause measurable changes in the voice pattern [51]. If voice analysis is to be used, analysis is likely to be performed by use of artificial neural networks as a natural extension of work that has already been performed for speech recognition [7,32,48].

2.4.3 Measures of Eye-Blink Activity. Eye blink activity, or electro-oculography (EOG) data, includes such measures as eye blink rate, duration of eye blinks, and eye blink latency relative to a stimuli. Endogenous eye blinks do not occur in response to specific environmental stimuli and have been found to vary as a function of the level of visual attention to a task [44,51]. Specifically, blink rate has been shown to decrease during times of high visual workload demand. Additionally, eye blinks tend to occur after a person takes in visual information, as may occur frequently in the rich and varied visual environment encountered by aircraft pilots during daytime flights. Overall, the use of eye blink data has demonstrated considerable utility with operator functions involving variation in the processing of visual information. This is evidenced in many studies where measures of eye-blink activity are among the most salient features used to discriminate between workload conditions [8,15,19,20,21,50,52]. But, EOG data does have its own inherent limitations and may be less satisfactory in tasks that involve significant manipulations to auditory or cognitive load, unassociated with visual stimuli [51].

2.4.4 Measures of Brain Activity. By placing an electrode on the scalp and using the appropriate electrolyte (an electrochemical gel), an on-going potential difference between the location on the skin and electrode can be recorded [29]. The resulting pattern of fluctuating microvolts can then be seen to exhibit wave characteristics. A plot of these voltage changes over time is called an electroencephalograph (EEG). These brainwave signals occur spontaneously and are the result of on-going electrical activity of the brain. The use of EEG data has been successfully applied to monitor workload in a number of multi-task environments [51]. Environments include both simulated and real-world driving of automobiles, air traffic control, and the piloting of aircraft. Several studies

have shown that EEG data can be used to help identify an operator's mental workload. Researchers with relevant published works include Brookings and Wilson [8], Caldwell, *et al.* [10], Hanskins and Wilson [26], Gevins, *et al.* [16,17,18], Greene *et al.* [19,21,22,23], Lizza [29], Morton and Wilson [31], Sirevaag, *et al.* [40], and Wilson and Eggemeier [51]. EEG normally includes all electrical activity observed at discrete locations on the scalp in the range of approximately 0.5 to 40 Hz. A maximum threshold of 40 Hz for EEG data is utilized as observed frequencies over 40 Hz are generally attributed to muscular activities.

Because any continuous "wave" function can be written as a linear combination of sinusoidal waves [24,28] by utilizing the theory of Fourier transforms, the average power observed from 0.5 to 40 Hz can be mapped into any number of smaller frequency bands. The Fourier transform provides a way of describing the time series EEG in terms of the frequency components of the signal. In the time domain, a continuous function can be represented by the form:

$$A^*(t) = \int a^*(t) e^{2\pi i \omega t} dt \qquad (2\text{-}13)$$

where a* is some quantity described as a function of time t. The signal is thus mapped from the time domain into a representation of amplitude A*. Amplitude can also be viewed as a function of frequency ω, and can be described by the following equation:

$$A^*(\omega) = \int a^*(\omega) e^{-2\pi i \omega t} dt \qquad (2\text{-}14)$$

where t is a finite period of time. If t is measured in seconds, then ω is measured in cycles per second or Hertz (Hz). Typically, the range of EEG data is decomposed into five distinct power bands. Table 2-1 defines the parameters of the frequency bands to be used in this research.

Table 2-1. Frequency Band Designations.

Band	Symbol	Frequency
Delta	Δ	0.5 - 3.0 Hertz
Theta	θ	4.0 - 7.0 Hertz
Alpha	α	8.0 - 12.0 Hertz
Beta	β	13.0 - 30.0 Hertz
UltraBeta	$\mu\beta$	31.0 - 42.0 Hertz

The average amplitude of power for a given band for each epoch can be determined using a more computationally efficient implementation of the Fourier transform. A Fast Fourier Transform (FFT) is used to perform this mapping by computing a discrete Fourier transform in a numerically efficient manner. For $2m$ data points, a FFT requires only $O(m\log_2 m)$ multiplications and $O(m\log_2 m)$ additions, where O represents a number "on the order of." This implementation saves considerably over the $(2m)^2$ multiplications plus $(2m)^2$ additions required to interpolate the triginometric polynomials of the Fourier transform by direct calculation [9].

Overall, research has shown that the use of EEG data can be used to discriminate between different mental workload requirements experienced by a subject. Additionally, research has shown that certain variances in specific power bands may be associated with mental workload changes of the operator. For example, studies have shown that α-band EEG activity normally decreases with an increase in cognitive demand [8,18,51], while θ-band EEG activity normally increases during periods of increased cognitive demand [10 ,18,26, 51]. Similarly, θ-band EEG activity has been shown to decrease from single to multiple tasks [51]. Finally, decreased operator performance has been associated with increases in α-band EEG activity [8] combined with decreases in θ-band and β-band EEG activity [51].

2.4.5 Measures of Hormone Levels. Of all the physiological features, measures of hormone levels are most difficult to identify for specific events. In response to sympatic nervous system stimulation, the adrenal glands release hormones into the blood system [51]. Evidence suggests that adrenaline levels are more influenced by mental effort and noradrenaline levels are more associated with physical effort. The amounts of these hormones can then be measured in a subjects blood, urine, or saliva level. Because of the requirement to take a physical sample from a subject, measurements of hormone levels are typically only taken after a relatively long period of time. Thus, severe limitations are placed on the ability to correlate hormone levels with specific events. Overall, the required samples and limited ability to correlate data with specific events make the use of hormone measurements less attractive than other physiological features for mental workload classification in an air traffic control or flight scenario.

2.4.6 Measures of Electrodermal Activity (EDA). By placing two electrodes on the skin and emitting a small current between the electrodes, a measure of impedance or resistance can be continuously observed over time. Thus, changes in the impedance on the skin can be detected. Variations in dermal impedance arise from sources such as perspiration level. One primary use of this measure is the modern lie detector, in which changes in electrodermal activity can normally be seen as distinguishable patterns between true and false responses given by a subject. Although the use of EDA data is relatively new for the assessment of mental workload in a multi-task environment, it provides an additional measure that is continuous in time and directly relates to a person's physiological state. This data can then be analyzed using saliency screening and

may prove to be useful in classifying mental workload levels that are associated with specific temporal events.

2.4.7 Summary of Psychophysiological Features. Overall, the use of multiple physiological features is required to adequately classify mental workload in a multiple task environment. Among those psychophysiological features described above, measures of heart rate, eye blink, respiration, EEG and EDA appear to be the least intrusive for use in work environments. These measures are capable of providing real time responses that can be used to assess mental workload at a specific instant in time. By their very nature, multi-task environments place demands on several aspects of a person's mental processing capabilities with no one measure able to adequately provide the necessary information to estimate operator workload [51]. Thus, by using data provided by multiple psychophysiological features, greater insight into the dynamics of pilot or air traffic controller mental workload can be obtained. In addition, more information about operator workload is possible by analyzing multiple features and their potential interactions, rather than looking at the separate measures individually.

III. Data Collection and Preprocessing

This chapter starts with a description of the experiment and the data collected by the Flight Psychophysiology Laboratory. Additionally, this chapter provides the methodology utilized to preprocess all of the psychophysiological data provided by AFRL/FPL. The preprocessing of the data includes the procedures used for transforming all "raw" psychophysiological data files into useful features that can be used for modeling efforts aimed at correctly identifying the workload level. Finally, some initial data "snooping" was performed on the preprocessed data. The primary purpose of the data "snooping" was to gain familiarity with the data and possibly achieve some insight as to how the different features are related to workload levels and to the other psychophysiological features. Specifically, many plots of the data were inspected to identify any visually obvious patterns, including but not limited to the identification of apparent outlying observations.

3.1 The MAT-B Experiment

The Flight Psychophysiology Laboratory of the Air Force Research Laboratory (AFRL/FPL) conducted experiments at Wright-Patterson AFB in 1998 and provided the data used for this effort. Twelve subjects participated in the experiment, with all participants completing an approximate hour-long scenario on two separate days. The experiment utilized the Multi-Attribute Task Battery (MAT-B). MAT-B was developed by NASA to support research efforts focused on human operator workload and strategic behavior. Specifically, the battery is user-interactive software that simulates workloads

analogous to tasks a flight crewmember would encounter [11]. In addition, MAT-B provides a high degree of experimenter control including the flexibility to predetermine the various levels of tasks to be presented to a test subject. In preparation for the two days of recorded scenarios, each subject trained on MAT-B on several days, until a consistent level of proficiency was achieved. This was performed to help reduce any potential effects of a learning curve and to allow subjects to achieve some familiarity with the battery. The experiment monitored and recorded both performance measures of required MAT-B tasks and psychophysiological features of each test subject.

MAT-B tasks include various monitoring, tracking, communication, and resource allocation responsibilities in a continually changing environment. Specifically, the monitoring task involves continuous visual observation where subjects respond to the absence of a green light, presence of a red light, or the movement of pointers away from specified midpoint levels. When out of specification, the system is either corrected by the subject or is automatically reset after a preset time interval. The tracking task involves using a joystick to keep a target centered within given limits, which simulates the demands of manual flight control. Additionally, during some of the lower workload levels the tracking task will provide an "AUTO" signal in which the tracking task is automated to simulate the reduced crewmember demands when utilizing an autopilot. The communication task simulates receiving auditory messages from air traffic control. In doing so, subjects must pay attention to messages received on a headset for his or her unique call sign and make frequency changes on the proper navigation or communication radio. Finally, the resource allocation task involves monitoring a simulated fuel system to maintain a specified level of fuel in two primary tanks. Subjects must react to

changing fuel levels and pump failures by switching on or off any of the eight different pumps to achieve the desired fuel levels. A representation of the MAT-B display created in *Microsoft Excel* can be seen in Figure 3-1.

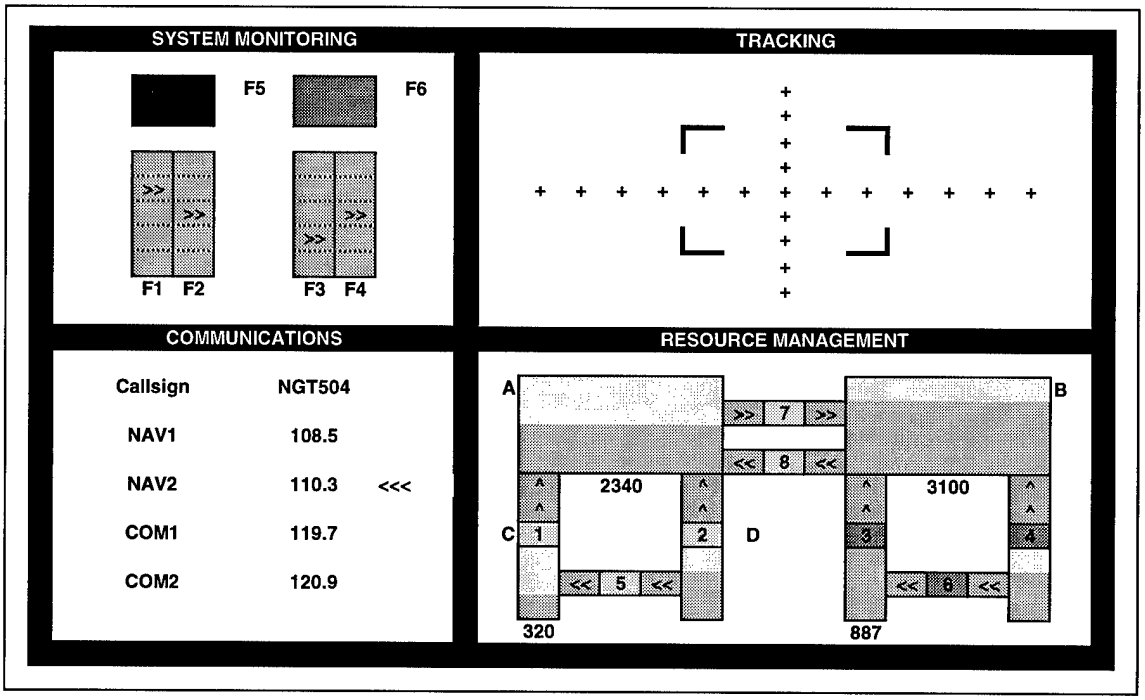

Figure 3-1. Sample MAT-B Display.

Three workload levels were introduced to the test subjects through the course of the experiment: low, medium, and overload. Low and medium levels offered tasks that could be completed by the test subject, while a subject could not complete all required tasks in an overload scenario. To create the distinct workload levels, the frequency of individual tasks for the subject to perform can be increased or decreased. Some specific changes to the subjects environment include varying the use of "AUTO" status for tracking, an increased number of system failures, more or less frequent calls from the air traffic control, and varying the probability of pump failure. Additionally, the level of tasks required of a subject in the low level were designed to be minimally taxing to

simulate a case of a crewmember utilizing an autopilot in a relatively low stress environment. In addition, the medium workload level was created to provide a work environment that is representative of a halfway point between low and overload.

The three workload levels were presented to each subject in approximate 15-minute "blocks," using one of six possible randomized orders designated A through F. Each block included 5-minutes of observations from each of the three distinct workload levels. The data files of each experimental block were recorded and designated by a subject number and a letter corresponding to the order of workload conditions. In addition, on each of the two days, subjects completed three of the 15-minute blocks. These blocks are identified as 1-6, with 1-3 corresponding to whether the block occurred 1st, 2nd, or 3rd on day 1, and 4-6 representing the three blocks in order on day 2. Thus, 09B2 represents subject 09 presented with workload order defined by sequence B as the 2nd of three 15-minute runs performed on day 1, as can be seen in Table 3-1.

Table 3-1. Experiment Workload Orders.

Label	Workload Order
A	Low - Medium - Overload
B	Low - Overload - Medium
C	Medium - Low - Overload
D	Medium - Overload - Low
E	Overload - Low - Medium
F	Overload - Medium - Low

Thus, approximately 45-minutes of data was collected in three different blocks, where the workloads were introduced in varied sequences. An example of a three blocks combination for one day is presented in Figure 3-2.

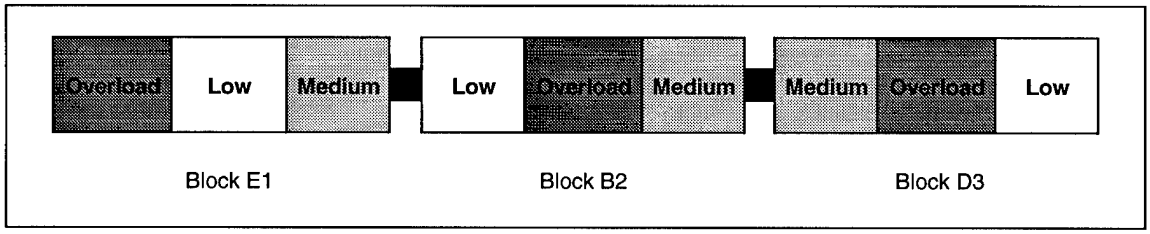

Figure 3-2. Sample Experiment Sequence.

Additionally, transition periods were included to facilitate change from one workload level to another. During this period, the tasks required of the test subject would gradually increase or decrease to the next level. Specifically, a 30-second transition period was included for a change of one workload level (e.g. low to medium or medium to overload), while a 60-second transition period was included for a change of two workload levels (e.g. overload down to low).

Thus, the data for all test subjects was collected in three distinct blocks each day which all contained equal periods of low, medium, and overload workload. These blocks can provide for a natural division into three separate groups, all with an equal sample of data from each of the three workload conditions. Additionally, this natural separation of observations will be useful for various modeling techniques, where one of the three blocks of data can be used as an independent validation set. More specifically, for ANN modeling these three blocks may quickly provide training, test, and validation sets of exemplars.

3.2 Psychophysiological Data Collected

Collected data included electroencephalography (EEG) electrode readings taken from 60 different locations on an individual's head on the first day and six locations on the second day. EEG data collection was performed differently on the two days to most

efficiently support multiple research objectives by the FPL. To facilitate possible day-to-day research efforts, the six matching locations from the first day are utilized (Figure 3-3). The location and naming of these sites are based on the International 10-20 system. The EEG location names first include a letter representation to designate the region of the brain. A number is then used to designate placement left or right of center, with odd numbers on the left, even on the right, and a "Z" for center locations. In the locations identified below, "O" represents the *ocular* region of the brain, where a majority of the processing of visual information is performed. Additionally, "F" designates the frontal region, "T" designates the two temporal regions, and "P" represents the parietal region (the middle division of the cerebral lobe).

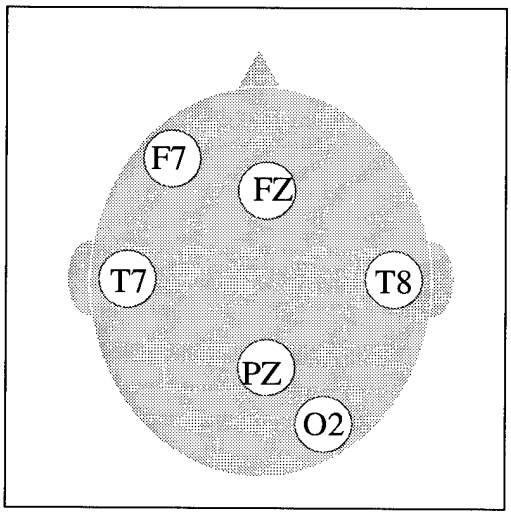

Figure 3-3. EEG Electrode Locations as Viewed from Top of Head.

Additionally, the EEG data was collected using two different hardware and software systems. *Neuroscan* was utilized to collect day 1 EEG data and includes the observed microvolts (μV) at each location sampled at 200 Hz. This EEG data is recorded and filtered in 5-minute workload periods. After collection of day 1 EEG signals,

Manscan V. 4.0 proprietary software was utilized by AFRL/FPL to remove undesired artifacts from the signals. Examples of removed noise includes effects from horizontal eye movement, vertical eye movement, AC interference, and muscle movements identified as observed potential above 15 μV at frequencies above 40Hz. In contrast, day 2 EEG data was collected using the *Workload Assessment Monitor (WAM)* that sampled the potential at six EEG locations at 128 Hz. Additionally, *WAM* performs its own proprietary filtering, provides EEG data after a FFT has been performed for each second, and provides a single data file for each 15-minute block.

In addition to the EEG data, raw physiological data for heart rate, eye-blinks, and respiration was collected identically on both days using *WAM,* and is recorded in 15-minute blocks. *WAM* provided data files including the time elapsed in milliseconds between discreet events including heartbeats, breaths, and eye-blinks. In addition, minimum and maximum amplitudes associated with each breath, the amplitude of each eye-blink, and the duration of each eye-blink were also provided. The final information provided is a list of event marks. The event marks contain the time in seconds from the start of an experimental block for the beginning and end of each of the three 5-minute periods of different workload levels.

3.3 EEG Processing

The following efforts were accomplished after the data was provided by AFRL/FPL. In order to use the EEG data, a series of steps were performed using code written in *Matlab* to process the "raw" EEG signals. Once processed the resulting features can be used by either linear or nonlinear models to classify workload level. As with any pattern recognition effort, preprocessing the data is a crucial step to facilitate the

most efficient use of the information available. Additionally, models can only be as accurate as the input will facilitate. Thus, input feature data processing was performed in a manner that has been shown effective by previous mental workload classification efforts.

As was previously mentioned, six EEG locations were selected to provide the input to form independent predictor variables. An example of one "raw" signal is provided in Figure 3-4.

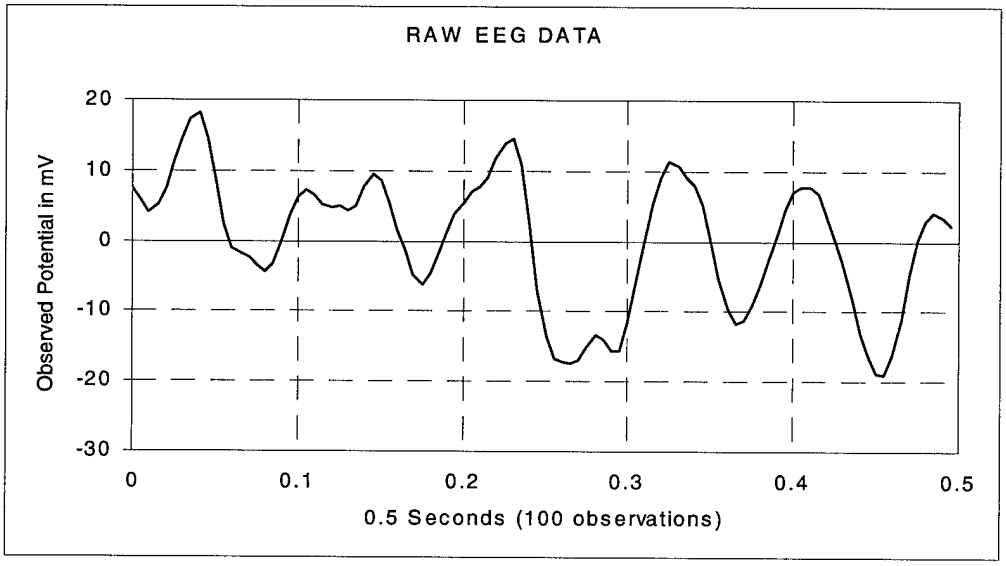

Figure 3-4. Raw EEG Signal from One Location over 0.5 Second.

This "raw" signal contains the combined power over a range of frequencies, with the amplitude of power being much greater for different frequencies. While it is not possible to determine the exact amplitude of any given frequency, two trends can be seen. First, the primary tend in the graph is a sinusoidal wave with peaks at approximately every 0.1 seconds. Thus, the majority of the power in this wave is probably contained in a frequency close to 10 Hz. Next, for this 0.5-second interval a downward trend is

apparent. If the trend were to increase over the next 0.5 second, this may correspond to significant power with a frequency of approximately 1 Hz.

To obtain an estimate of the power for frequencies of interest, the first step of processing day 1 EEG data is to perform a Fast Fourier Transform (FFT) of the raw signal from all six locations. To match day 2 data, an FFT was performed on each EEG signal for every 1-second of raw data. Additionally, according to the Nyquist sampling theorem, estimates for power can only be made for frequencies up to $f_S/2$, where f_S is the sampling frequency [30]. Thus, with a sampling frequency of 200 Hz, a 100-point FFT with 1-Hz resolution was performed using *Matlab* to obtain power estimates for frequencies of 1 to 100Hz. An example of the power estimates by frequency over a 1-second window is presented in Figure 3-5, and is known as a periodogram [30].

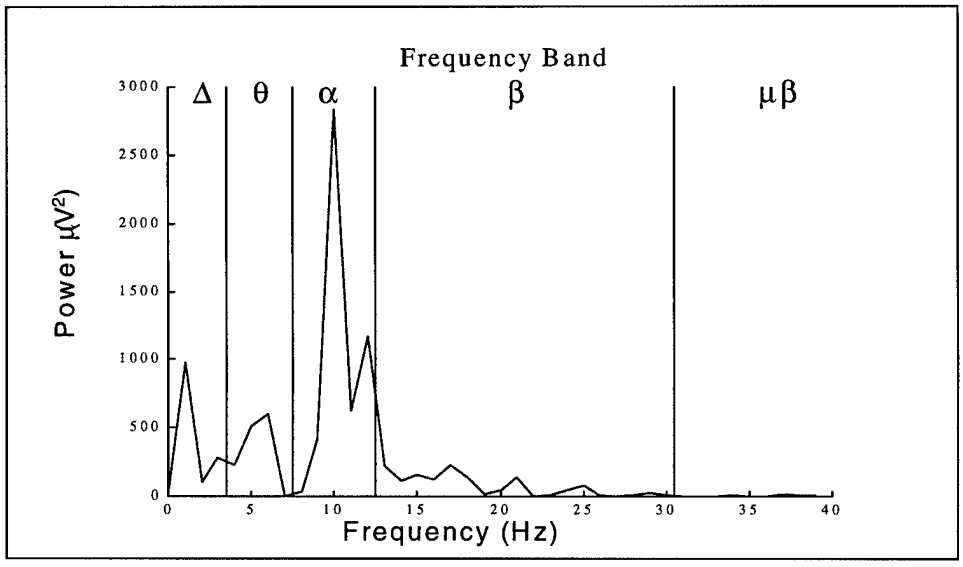

Figure 3-5. Fast Fourier Transform of One EEG Electrode (Periodogram).

As seen in Figure 3-5, the greatest amplitude of power occurs at about 10 Hz, while a significant power spike can also be seen down at 1 Hz.

While a 100-point FFT was calculated, only power estimates for frequencies of 1 to 40 Hz are realized. This was expected, as the observed potential was filtered by AFRL/FPL using *Manscan V. 4.0* software to remove frequencies below 0.5 Hz and those above 40 Hz. The y-axis of the above figure is now power, expressed in microvolts^2, which was obtained by multiplying the frequency-decomposed transformed signal by its complex conjugate. The x-axis is the frequency in Hz, with the vertical lines within the plot representing the boundaries between the five frequency bandwidths associated with EEG signals. These bandwidths are the same as presented in Chapter 2 as Table 2-2.

After taking the 1-second FFT, the average power within each frequency band is then summed to produce a power estimate for each of the five frequency bandwidths. This process effectively acts as a bank of five elliptical filters. Each frequency bands' lower and upper frequencies represent the cutoff frequencies defining the passband of an elliptical filter, and the accuracy of the FFT algorithm determines the effective ripple allowed in the passband and the attenuation of power in the stopband [30]. With each EEG signal decomposed into power estimates in five bandwidths, visual inspection of the data indicates considerable noise that may have more desirable, less noisy, underlying trends. Figure 3-6 provides a plot of a 5-minute window of the average potential in the five frequency bands taken at each second. In contrast to the power estimates of theta, alpha, and beta frequency bands that overlap, the power in the ultrabeta band is significantly less. This is apparent in Figure 3-6, and is expected. From the periodogram presented as Figure 3-5, the estimate of power for frequencies in the ultrabeta frequency range is considerably less than power estimates in the lower frequencies.

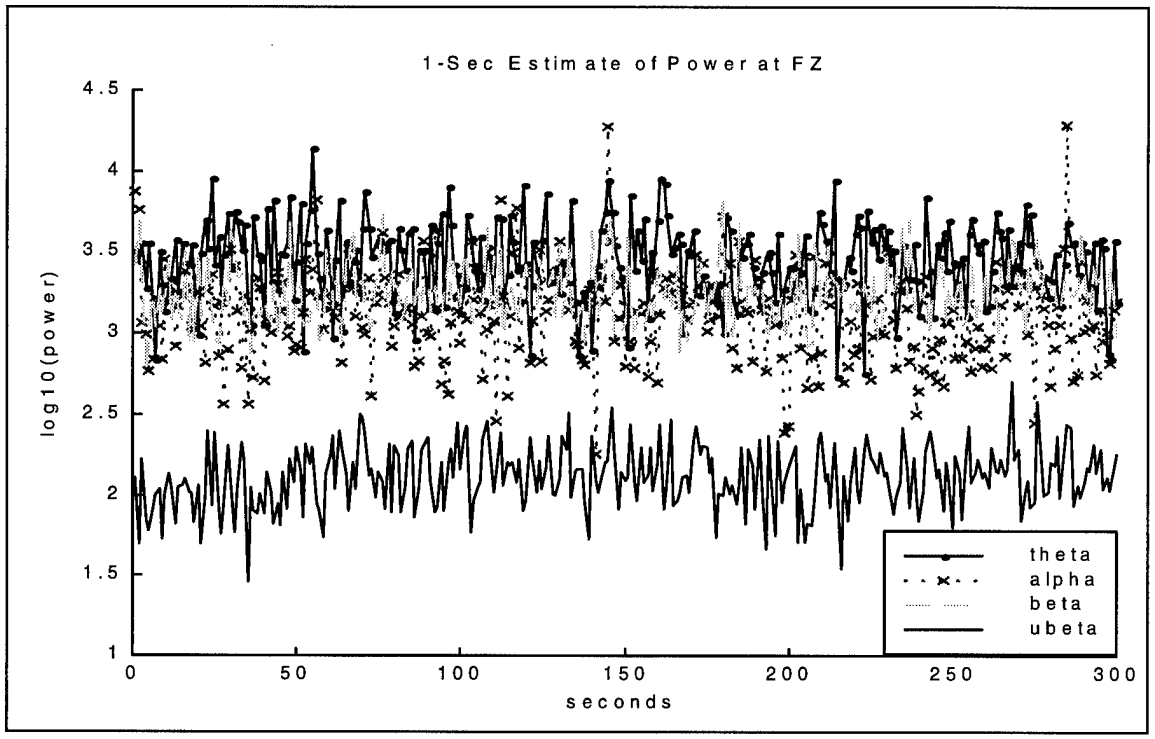

Figure 3-6. Log10 of Average Potential in 4 Frequency Bands.

Unfortunately, periodogram estimates of power obtained from FFT decomposition often have large variance, which do not decrease even if sample size is increased [30]. Therefore, in order to smooth the EEG power observations, all power estimates were averaged over a 10-second window including 5-seconds of overlap with the previous observation (Figure 3-7).

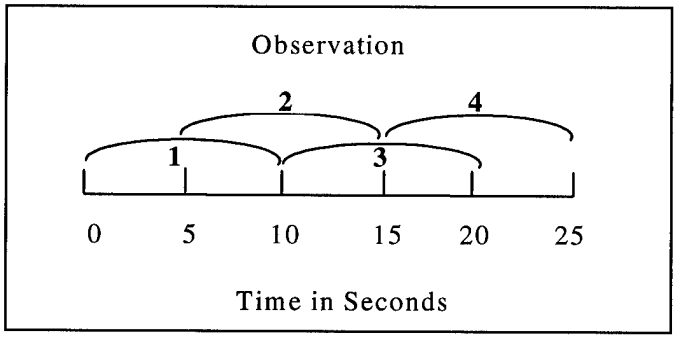

Figure 3-7. Data Sampling Overlap.

This technique has been demonstrated to be effective at reducing noise, with recent utilization by Greene [19], Greene et. al [20,23], and Russel et. al. [38,39] in similar mental workload classification efforts utilizing ANNs. Thus, starting with 300 1-second power estimates, 30 non-overlapping 10-second windows can be formed with an additional 29 overlapping windows. The net result is 59 exemplars for a 5-minute period and 177 exemplars over a 15-minute block of time.

Finally, after the 10-second averages have been computed, the $\log_{10}$ of the average power is taken. An example of a fully processed 5-minute block of EEG data is presented in Figure 3-8. The y-axis is scaled as $\log_{10}$ of average microvolts^2 (μV^2), for each bandwith, and the x-axis is scaled in seconds.

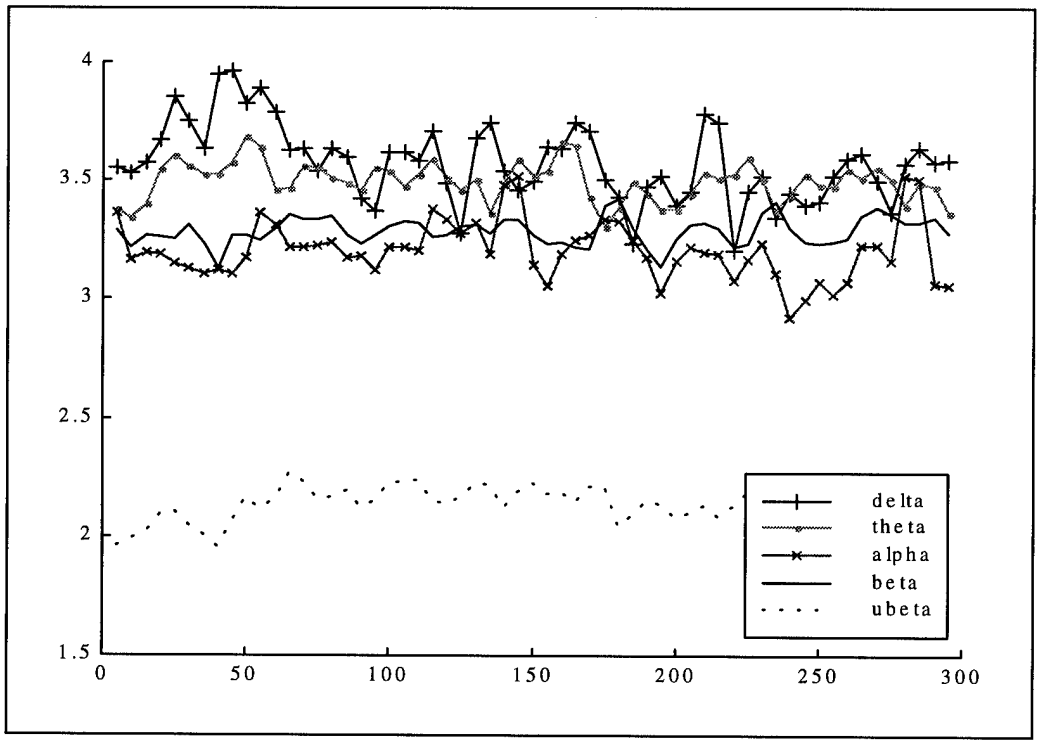

Figure 3-8. Processed EEG Signal with 5 Second Overlap (5 minutes).

Thus, with six locations and five bandwidths, 30 different EEG variables are created and can be labeled by the placement of the reading and the bandwidth of the frequency. A summary of the steps used to process the day 1 EEG signals is provided as Figure 3-9.

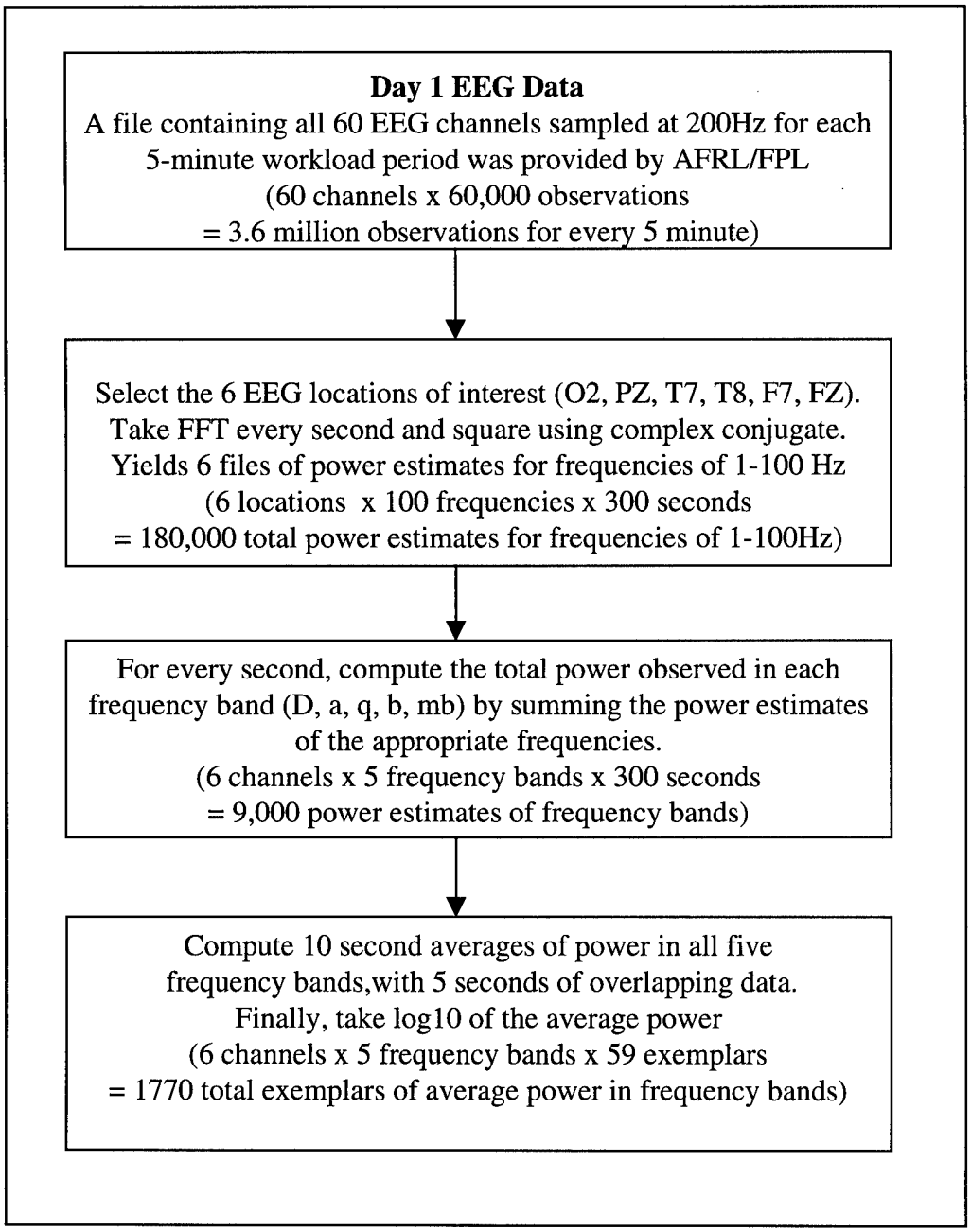

Figure 3-9. EEG Data Processing.

3.4 Physiological Feature Processing

As stated in section 3.1, data files including the time of discrete physiological events including heartbeats, eye-blinks and the number of breaths taken were available for all 15-minute blocks of time. Additionally, an event file was provided including the start and stop of the low, medium, and overload workload periods. The following three subsections describe the procedure used to process the raw physiological data provided by AFRL/FPL into six distinct physiological features.

3.4.1 Electrocardiography (ECG). Two ECG features were processed and provide a measure of the cardiovascular activity in terms of rate per unit time and change in rate during a unit of time. To match EEG features, an average heart rate was calculated for each 10-second window. This was accomplished in *Matlab* by identifying all observed beats within a given 10-second window and calculating the average interval observed between two adjacent beats. The average interval between beats in milliseconds was then transformed into beats per minute by inverting the average time between beats and multiplying by 60,000 milliseconds per minute. A sample containing 15-minutes of processed heart rate is included as Figure 3-10.

To obtain a measure of heart rate variability during any 10-second window a first order polynomial was fit using ordinary least squares to all observed time intervals between heartbeats in a window. The slope term of the least squares polynomial was then used as an estimate of the change in heart rate. Because the inter-beat intervals (IBI), are expressed in seconds, the slope is in units representing the change in IBI seconds per 10-second window. Finally, since the magnitude of change is an adequate estimator of variability during any 10-second window, the absolute value of the slope of

the heart rate was taken. An example of processed heart rate variability is provided as Figure 3-11.

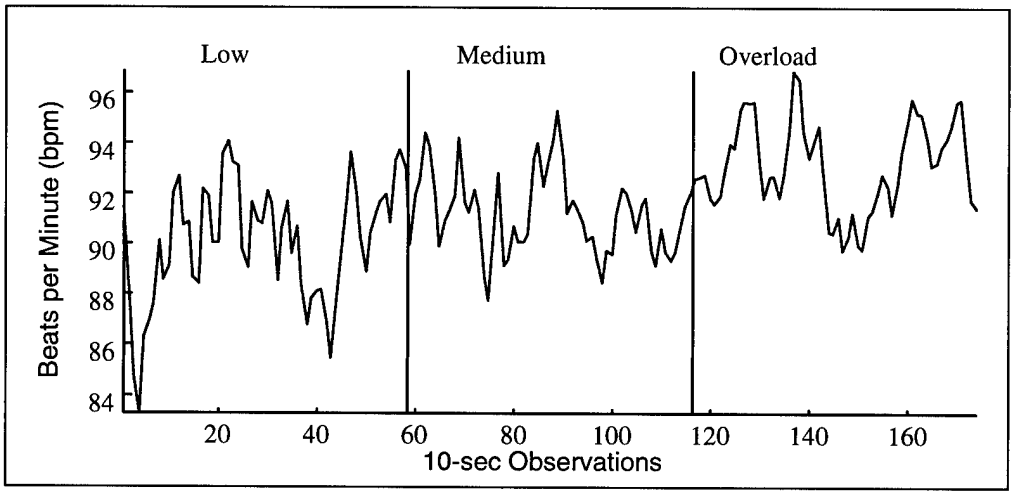

Figure 3-10. Heart Rate.

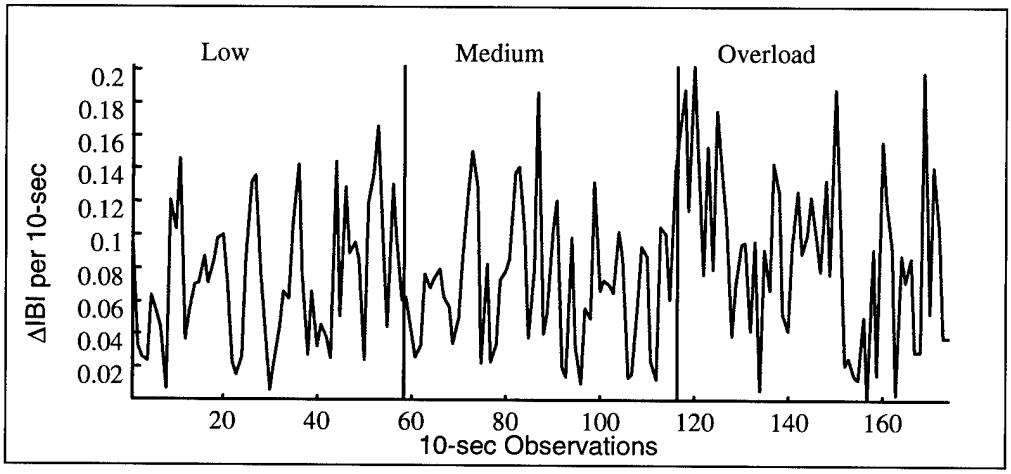

Figure 3-11. Variance of Heart Rate.

A summary of the steps taken to process the ECG data is provided as Figure 3-12.

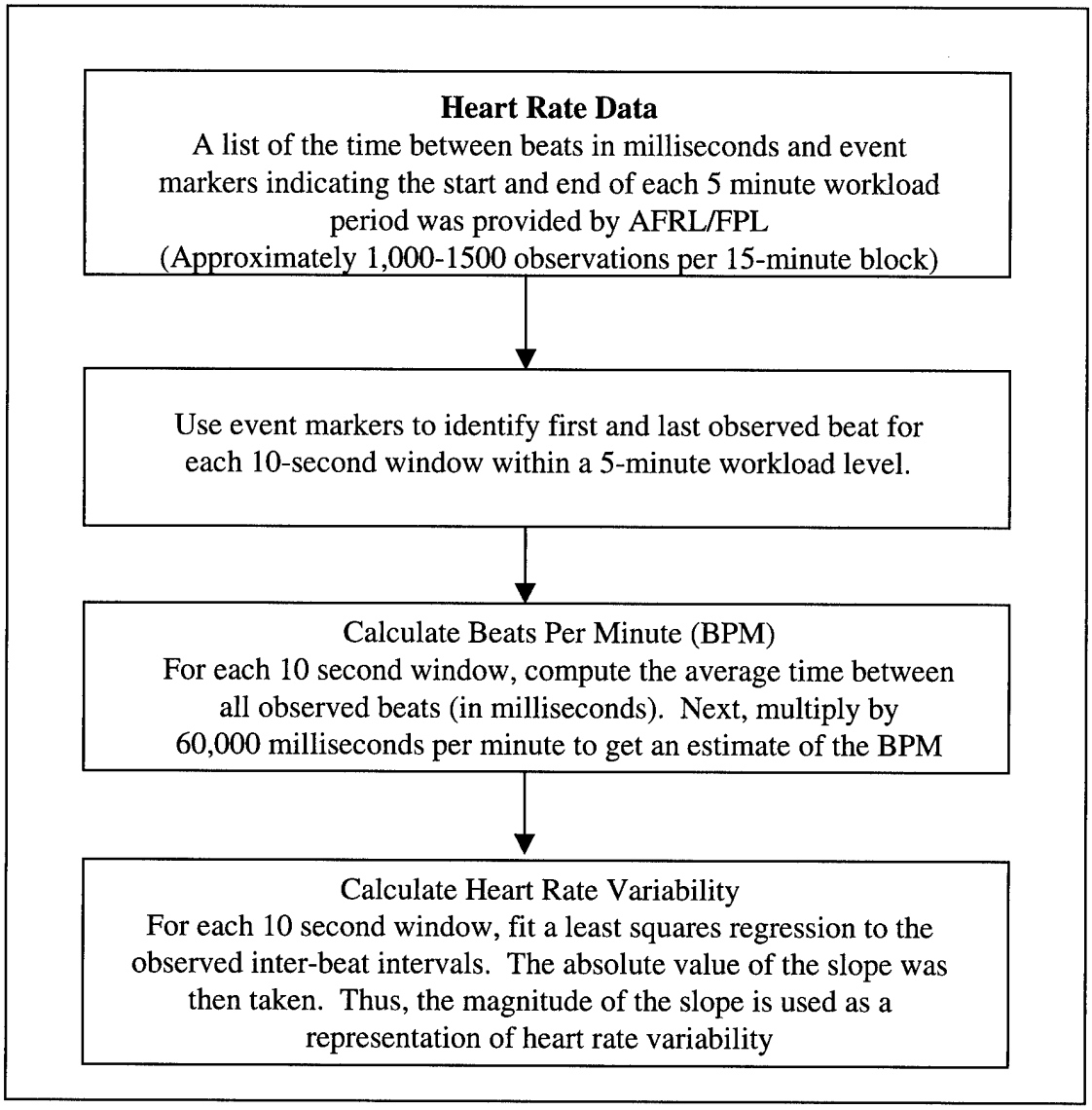

Figure 3-12. Heart Rate Data Processing.

3.4.2 Electro-oculographyy (EOG). Two EOG features were processed and provide a measure of eye movement in terms of the discrete number of blinks per unit of time and as the average time between blinks. Again, features were calculated for 10-second windows by using code written in *Matlab.* To calculate the number of blinks, all blinks within a window were simply identified and counted. Typical values for a 10-second window ranged from no blinks up to six or more blinks, as can be seen in Figure 3-13.

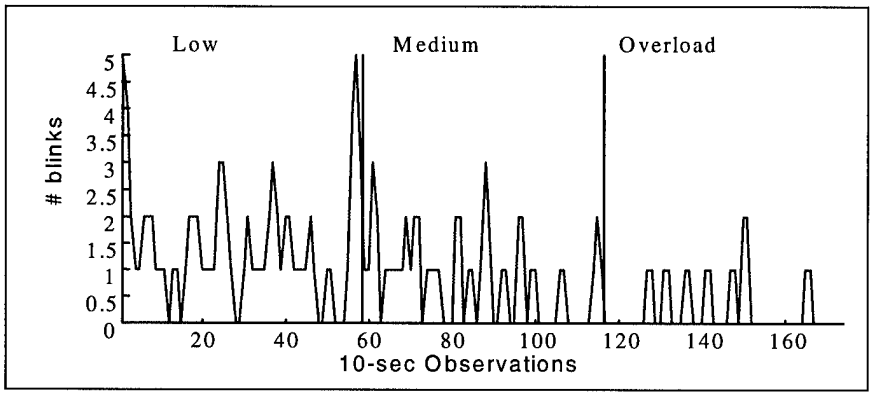

Figure 3-13. Observed Eye-Blinks.

Calculating the average time between blinks for a given window is not as straight forward and includes three potential scenarios. First, if two or more blinks occurred, an average time was calculated for all observed inter-blink intervals (IBLIs), as was done for heart rate. Next, if only one blink occurred during the 10-second window, the prior blink was found and the time between the two blinks was used. Finally, if no blinks occurred, the time the last blink occurred was subtracted from the time at the end of the current window. Thus, if no blinks occurred the value assigned to IBLI is the time the subject has gone without blinking. A sample plot of the IBLIs is presented as Figure 3-14, and the process is summarized in Figure 3-15.

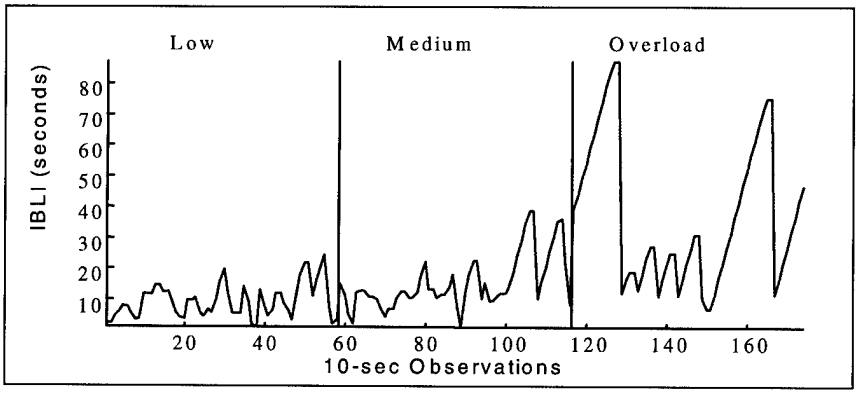

Figure 3-14. Average Time Between Blinks.

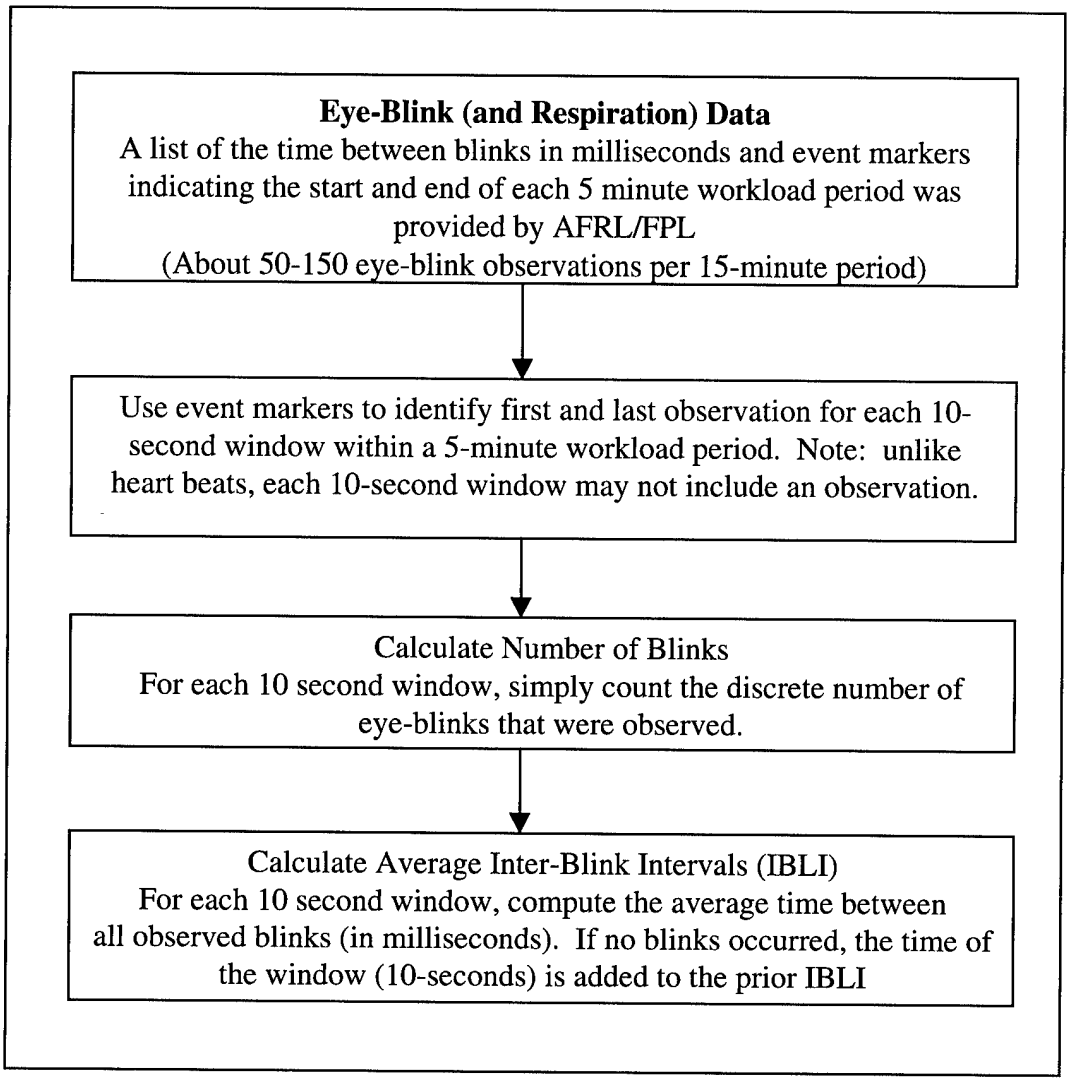

Figure 3-15. Eye-Blink Data Processing.

3.4.3 Respiration. Two respiration features were processed and provide a discrete measure of the number of breaths per unit of time and a measure of the average time between breaths. These two features were processed identically to the eye-blink features as described in Section 3.4.2. Most values for the number of breaths in a 10-second window ranged anywhere between zero to six, with most 10-second windows containing between one and four observed breaths. A representative plot for the number of breaths observed in a 15-minute period is included as Figure 3-16, and a representative plot of the

average inter-breath interval (IBRI) is provided as Figure 3-17. Figure 3-15 may be referenced as a summary of the steps taken to process the respiration data.

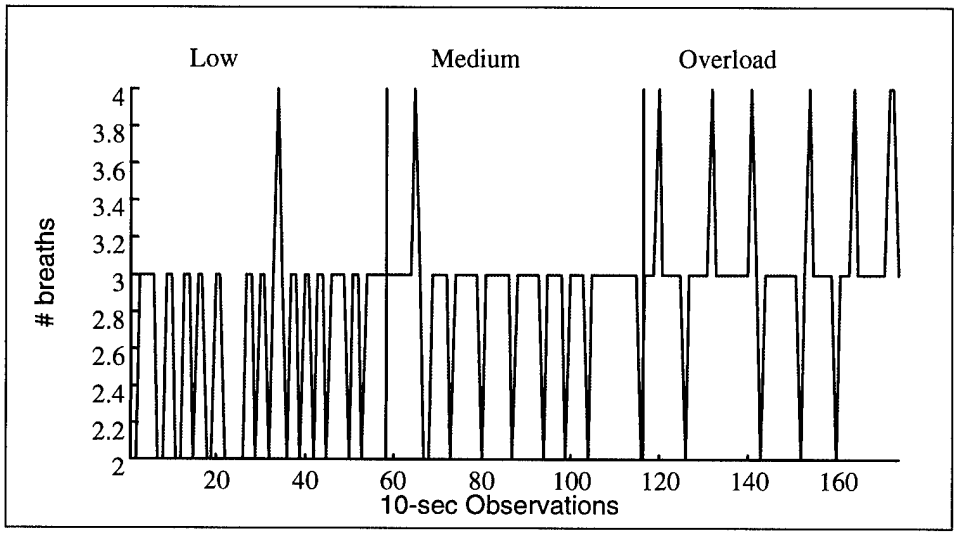

Figure 3-16. Observed Breaths.

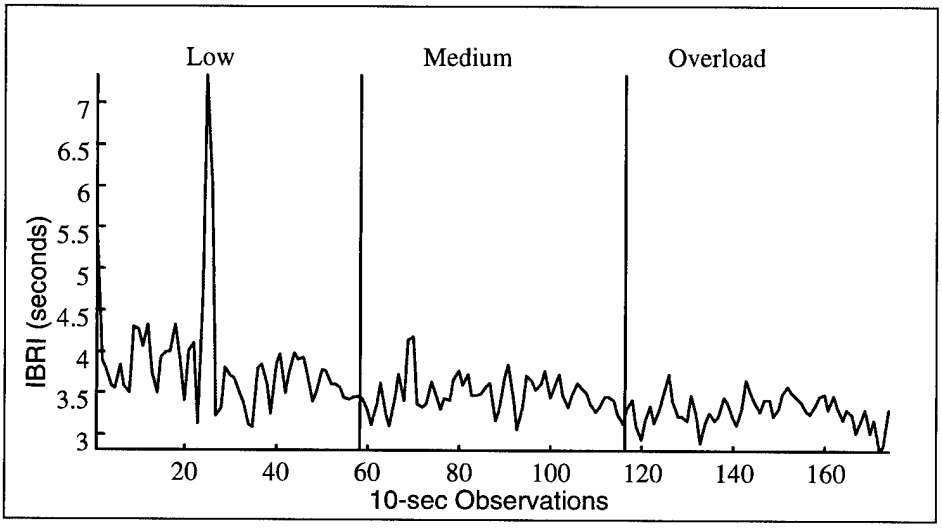

Figure 3-17. Average Time Between Breaths.

3.5 *Summary of Features*

After all data preprocessing was completed, a total of 36 psychophysiological features were available for use to discriminate between the three various levels of tasks presented in the simulated flight environment. Table 3-2 lists the 36 features processed for each individual from the experiment including feature row number, abbreviated name, and units. In addition, to facilitate ANN saliency screening of features using the SNR measure from Equation 2-13, a 37th feature of random noise, uniformly distributed from 0.0 to 1.0, was added. Features 38, 39, and 40 were then added as an indicator vector to identify the true workload level for any particular observation. These features can be used to calculate the classification accuracy of any given model, in addition to being utilized as the desired training targets for an MLP. Feature 38 was set to 0.9 if the true workload was low, feature 39 was set to 0.9 if the true workload was medium, and feature 40 was set to 0.9 if the true workload was the overload condition. For levels not specified as above, the values were set to 0.1.

Table 3-2. Database Variables.

Feature #	Name	Description	Units
1	O2-d	Power in Δ Band at O2	$Log_{10}(\mu V^2)$
2	O2-t	Power in θ Band at O2	$Log_{10}(\mu V^2)$
3	O2-a	Power in α Band at O2	$Log_{10}(\mu V^2)$
4	O2-b	Power in β Band at O2	$Log_{10}(\mu V^2)$
5	O2-ub	Power in μβ Band at O2	$Log_{10}(\mu V^2)$
6	PZ-d	Power in Δ Band at PZ	$Log_{10}(\mu V^2)$
7	PZ-t	Power in θ Band at PZ	$Log_{10}(\mu V^2)$
8	PZ-a	Power in α Band at PZ	$Log_{10}(\mu V^2)$
9	PZ-b	Power in β Band at PZ	$Log_{10}(\mu V^2)$
10	PZ-ub	Power in μβ Band at PZ	$Log_{10}(\mu V^2)$
11	T7-d	Power in Δ Band at T7	$Log_{10}(\mu V^2)$
12	T7-t	Power in θ Band at T7	$Log_{10}(\mu V^2)$
13	T7-a	Power in α Band at T7	$Log_{10}(\mu V^2)$
14	T7-b	Power in β Band at T7	$Log_{10}(\mu V^2)$
15	T7-ub	Power in μβ Band at T7	$Log_{10}(\mu V^2)$

16	T8-d	Power in Δ Band at T8	Log_{10} (μV^2)
17	T8-t	Power in θ Band at T8	Log_{10} (μV^2)
18	T8-a	Power in α Band at T8	Log_{10} (μV^2)
19	T8-b	Power in β Band at T8	Log_{10} (μV^2)
20	T8-ub	Power in μβ Band at T8	Log_{10} (μV^2)
21	F7-d	Power in Δ Band at F7	Log_{10} (μV^2)
22	F7-t	Power in θ Band at F7	Log_{10} (μV^2)
23	F7-a	Power in α Band at F7	Log_{10} (μV^2)
24	F7-b	Power in β Band at F7	Log_{10} (μV^2)
25	F7-ub	Power in μβ Band at F7	Log_{10} (μV^2)
26	FZ-d	Power in Δ Band at FZ	Log_{10} (μV^2)
27	FZ-t	Power in θ Band at FZ	Log_{10} (μV^2)
28	FZ-a	Power in α Band at FZ	Log_{10} (μV^2)
29	FZ-b	Power in β Band at FZ	Log_{10} (μV^2)
30	FZ-ub	Power in μβ Band at FZ	Log_{10} (μV^2)
31	HR	Heart Rate	bpm
32	HrVar	Heart Rate Variability	Δ sec per 10-sec
33	blnks	Number of Eye-Blinks	# blinks per 10-sec
34	IBLI	Eye-Blink Interval	seconds
35	brths	Number of Breaths	# breaths per 10-sec
36	IBRI	Breath Interval	seconds
37	Noise	Random Uniform(0,1)	none
38	Low	(0.9. if Low, 0.1 o.w.)	none
39	Medium	(0.9. if Medium, 0.1 o.w.)	none
40	Overload	(0.9. if Overload, 0.1 o.w.)	none

3.6 Initial Data Inspection

An initial effort was made to become familiar with the data. In addition to reviewing plots of all 36 features similar to those presented in the prior sections, *SAS-JMP* statistical software was utilized to examine correlations and to identify potential outlying data points. A sample of a correlation matrix of all ultrabeta features for one day (531 observations) is presented as Table 3-3. Of significance in this particular matrix are the high levels of correlation between some of the features as indicated by bold print. Additionally, while O2, PZ, T7, T8, and FZ are all positively correlated, F7 appears to behave opposite to these features as indicated by negative correlations. This observation is quite apparent when examining the scatterplot of all correlations in Figure 3-18.

Additionally, Figure 3-18. includes data points that appear as potential outlying, non-representative data points. To further investigate this possibility, the Mahalanobis distances of all observations were calculated. The Mahalanobis distance is denoted by D^2 and is calculated as:

$$D_i^2 = (\mathbf{x}_i - \mathbf{x}_{ave})'\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_{ave}) \quad (3\text{-}1)$$

where $\mathbf{x}_i$ is the vector of values at observation i, $\mathbf{x}_{ave}$ is the sample mean, and $\mathbf{S}$ is the sample covariance matrix. Unlike Euclidian distance statistics, the Mahalanobis distance explicitly accounts for correlations between variables [14]. Figure 3-19 is a plot of D^2 values for ultrabeta observations which reveals a pattern of outlying data points.

Table 3-3. SAS-JMP ultrabeta Correlation Matrix.

Correlations						
Variable	O2-ub	PZ-ub	T7-ub	T8-ub	F7-ub	FZ-ub
O2-ub	**1**	**0.911**	0.648	**0.791**	-0.662	0.758
PZ-ub	**0.911**	**1**	0.684	**0.792**	-0.630	**0.853**
T7-ub	0.648	0.684	**1**	0.636	-0.295	0.582
T8-ub	**0.791**	**0.792**	0.636	**1**	-0.683	0.649
F7-ub	-0.662	-0.630	-0.295	-0.683	**1**	-0.433
FZ-ub	0.758	**0.853**	0.582	0.649	-0.433	**1**

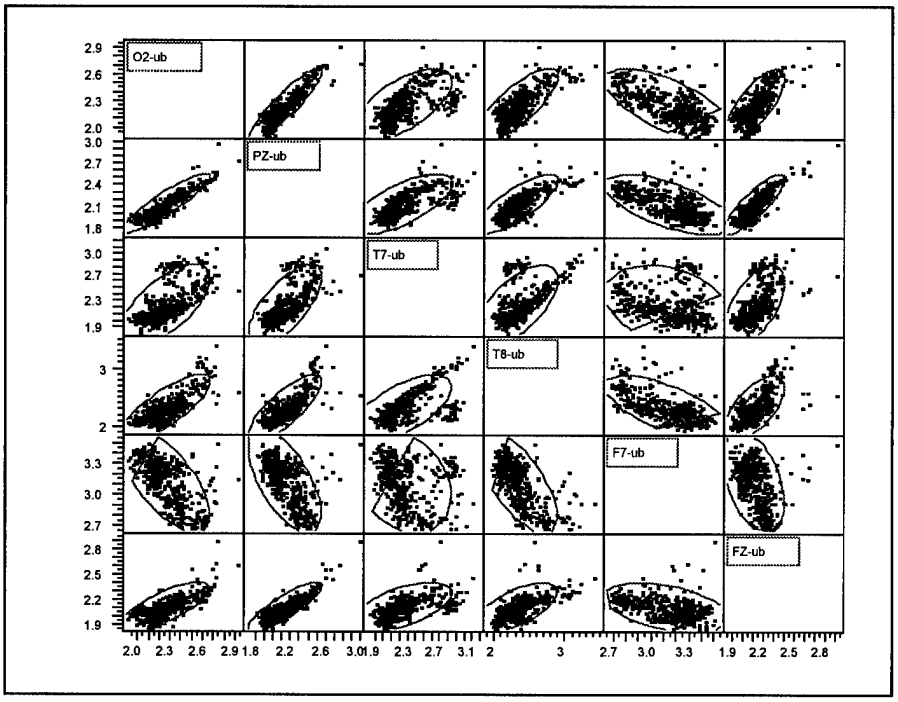

Figure 3-18. SAS-JMP Scatterplot of ultrabeta Correlations.

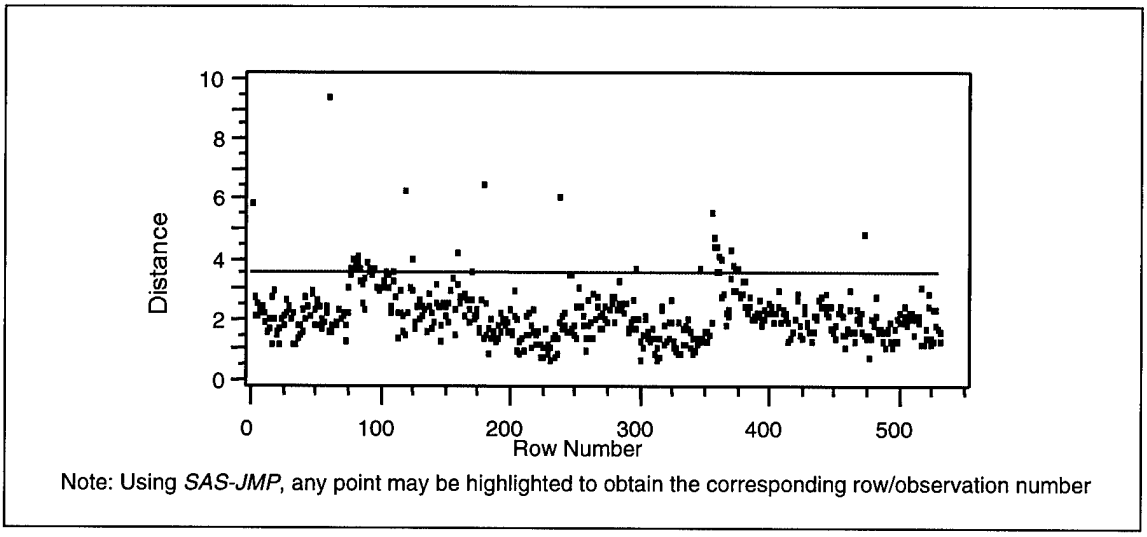

Figure 3-19. Plot of Mahalanobis Distances for ultrabeta Features.

From the above plot, observations with Mahalanobis distance greater than 5 were identified as observations #1, #60, #119, #178, #237, #355, and possibly #473. These observations present a specific pattern of interest, with each of these identified

outliers corresponding to an observation number of (59 • n)+1, where n is an integer. Thus, the identified outlying observations are the 1st calculated observation in each 5-minute workload condition. Because the subjects were gradually transitioned between levels, physiological arguments do not support the observed spikes in the ultrabeta power. These outliers are likely to be attributable to data recording or the data processing technique utilized. Consequently, the 1st 10-second window of data will not be used as it does not appear representative of the population.

To support the removal of these outliers, plots of the EEG data with and without the first observation are included. Figure 3-20 includes all processed EEG data from one electrode arranged by 5-minute periods of low, medium and overload workload. This plot clearly shows the first observation of each workload period as being abnormally large. Figure 3-21 presents the same data without the first observation, in which the ultrabeta observations appear to behave much more consistently.

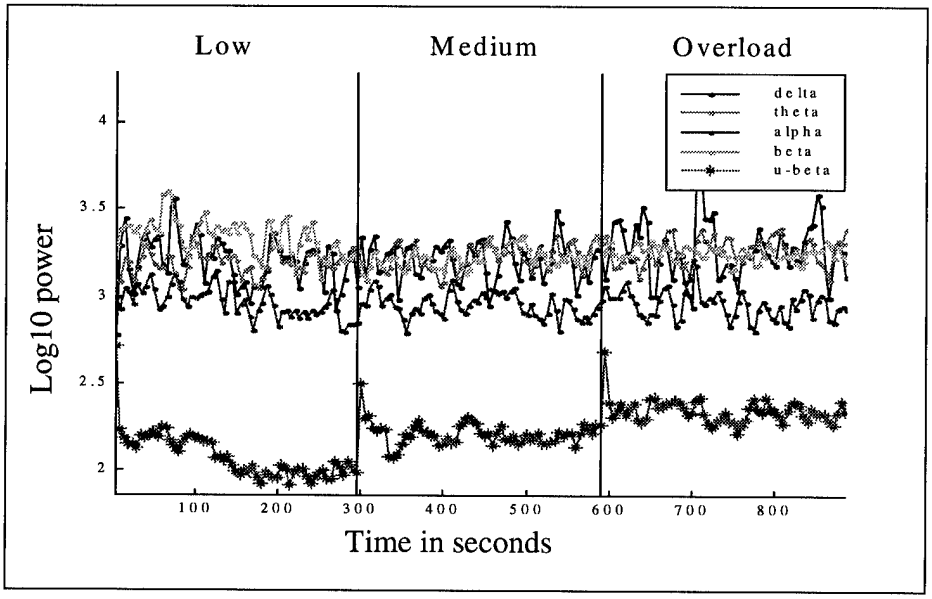

Figure 3-20. EEG Data with All Exemplars.

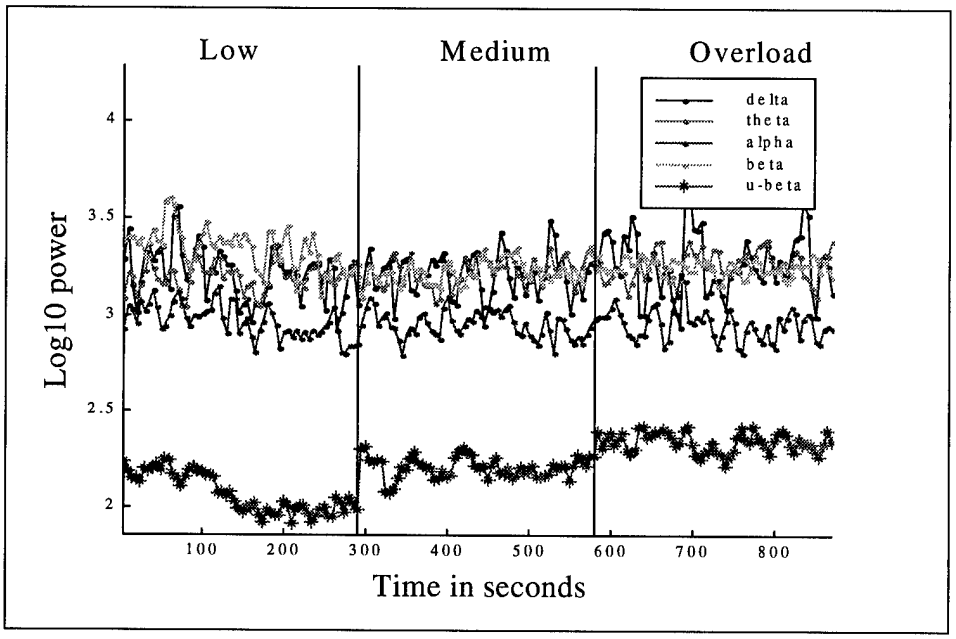

Figure 3-21. EEG Data with First Exemplar Removed.

3.7 Data Preprocessing Findings and Summary

After visual inspection of many plots of data, many trends appear which may prove fruitful in devising a model to accurately identify workload levels. Specifically, visual inspection of EEG data show that in many cases observed power in the ultrabeta frequency band at O2 and PZ appears to increase with increased workload. Also noticeable are trends with the eye-blink, respiration, and to a lesser extent, the heart-rate data. As is physiologically predicted, the number of eye-blink tends to decrease as visual tasks become more demanding [8,15,19,20,21,50,52], the number of breaths tends to increase with increasing workload [8,50,51], and heart-rates tend to increase with more demanding work levels [18,27,35,50,51,52].

As a final note, data from a total of eight subjects was processed and will be used in the research effort. To maintain consistency with the AFRL/FPL, the subject numbers were retained as provided from FPLwith the raw data. These numbers were assigned in

order as subjects volunteered to be part of this MAT-B workload experiment. Unfortunately some subjects did not complete all training and testing for the experiment. Thus, subjects with numbers from 02 to 16 are used in this effort. All eight test subjects by number including their specific workload patterns are identified in Table 3-4 below.

Table 3-4. Workload Presentation by Subject.

Subject	Block 1	Block 2	Block 3
02	C = (Me-Lo-Ov)	E = (Ov-Lo-Me)	B = (Lo-Ov-Me)
03	E = (Ov-Lo-Me)	B = (Lo-Ov-Me)	D = (Me-Ov-Lo)
04	B = (Lo-Ov-Me)	D = (Me-Ov-Lo)	F = (Ov-Me-Lo)
05	D = (Me-Ov-Lo)	F = (Ov-Me-Lo)	A = (Lo-Me-Ov)
09	E = (Ov-Lo-Me)	B = (Lo-Ov-Me)	D = (Me-Ov-Lo)
11	D = (Me-Ov-Lo)	F = (Ov-Me-Lo)	A = (Lo-Me-Ov)
13	A = (Lo-Me-Ov)	C = (Me-Lo-Ov)	E = (Ov-Lo-Me)
16	B = (Lo-Ov-Me)	D = (Me-Ov-Lo)	F = (Ov-Me-Lo)

IV. Methodology

This chapter includes a description of the methodologies utilized to classify workload using the processed psychophysiological features described in Chapter 3. Specifically, sections are devoted to the initial modeling efforts, the use of discriminant analysis for feature selection and workload classification for individuals, and the use of MLP ANNs for feature selection and workload classification of individuals. Finally, the methodology to assess the feasibility of a single MLP ANN is presented. This feasibility assessment includes both discriminant analysis and MLP ANNs to determine a set of salient group features and some initial ANN models to assess workload classification for the group as a whole.

4.1 Initial Modeling Efforts

To assess feasibility and to gain insight as to how well the data could be classified using the linear and non-linear models, subject 09 was used as a pathfinder for some initial classification efforts. Subject 09 was selected simply because this was the first data available from the Flight Psychophysiology Laboratory. Following are the methodologies and results of the first attempts at classification using a two-class discriminant model and a three-class MLP ANN model. In addition, from these pathfinder efforts, additional steps for data preprocessing and ANN parameter settings were determined before continuing efforts aimed toward determining if "one net could fit all."

4.1.1 Initial Two-Class Discriminant Model. A description of multivariate discriminant analysis was presented in Chapter 2 where group classification of an observation was assigned to the population with the greatest associated probability. While using scores based on probability is suitable for any number of classes, it is somewhat difficult to visualize the discriminant function (d_k^Q) scores. To facilitate visualization and to gain insight into the input data, Fisher's two-class model was first utilized. Fisher's two class model determines one discriminant score for each observation. For a sample of m observations with n input features per observation, this score is determined as:

$$\mathbf{Y} = \mathbf{b}'\mathbf{X} \tag{4-1}$$

where $\mathbf{Y}$ is a vector of m discriminant scores, $\mathbf{b}$ is a vector of n discriminant weights, and $\mathbf{X}$ is an $n \times m$ matrix containing all the observed independent input features. For this application, n = 36 psychophysiological features or input variables and m = the total number of 10-second observations. If the covariance structures of the two populations are assumed to be statistically equivalent, the vector of discriminant weights can be obtained as:

$$\mathbf{b} = \mathbf{S}^{-1}(\mathbf{x}_{1ave}-\mathbf{x}_{2ave}) \tag{4-2}$$

where $\mathbf{S}^{-1}$ is the inverse of the estimated pooled covariance matrix, $\mathbf{x}_{1ave}$ is the vector of average input features for population 1, and $\mathbf{x}_{2ave}$ is the vector of average input features for population 2. An optimal separation boundary is then determined to differentiate the two-classes. Typically, the separation point is taken as the midpoint between the average discriminant scores for the two populations, although with an unequal number of samples in each of the two classes, the optimal point of separation ($\mathbf{Y}_c$*) becomes:

$$\mathbf{Y}_c^* = \frac{n_2\mathbf{Y}_{1ave} + n_1\mathbf{Y}_{2ave}}{n_1 + n_2} \tag{4-3}$$

where $\mathbf{Y}_{1ave}$ and $\mathbf{Y}_{2ave}$ are the average discriminant scores from populations 1 and 2, and n_1 and n_2 are the number of observations in their respective populations.

Before discriminant analysis was attempted, one additional preprocessing step was performed. Because the data was processed with overlapping 5-second intervals, each observation was clearly not independent from the two other observations that also included some of the same "raw" data. Thus, for this first attempt at discriminant analysis and all future linear efforts, every other observation is not used. For example, only observations 1, 3, 5, 7, and 9 would be used from the set containing the first 10 observations. A plot of the discriminant scores in time order is provided as Figure 4-1.

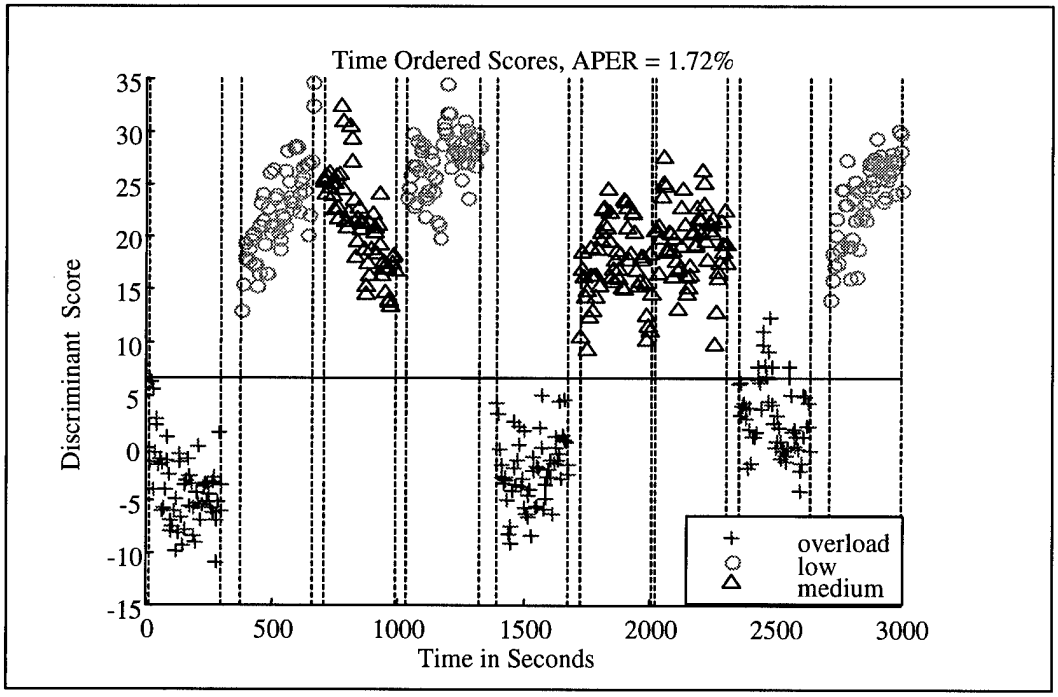

Figure 4-1. Time Ordered Discriminant Scores.

The model used to produce the plot in Figure 4-1 utilized all 36 psychophysiological features and all observations to define the model parameters. Thus, only an apparent error rate (APER) can be calculated. Additionally, both low and medium workloads were combined to create a single "nominal" workload condition. This provided for a two-group classification problem of nominal vs. overload conditions.

For this effort the APER was about 1.7%. In both Figure 4-1 and Figure 4-2, the discriminant scores are the y-variables, with the optimal separation point Y_c^* identified by a solid horizontal line through the middle of the graphs. In addition, dashed vertical lines represent the boundaries of each observed workload level. A distinct symbol is used for each observation representing it's true workload class. From Figure 4-1, most of the misclassifications can be seen in the last overload period where actual overload observation discriminant scores (+'s) are above the separation line and are misclassified as nominal workload. Overall, with the small APER, this effort shows that this nominal vs. overload problem is linearly separable. Thus, after starting with a simple model and two-class problem, initial workload classification feasibility has been demonstrated. A more complex classification problem including the three classes will be attempted in future work.

Also, from Figure 4-1, definite trends in the discriminant scores are evident even after the overlapping data points were removed. This is seen as the discriminant scores of low workload increases and drifts away from the separation line with no chance of misclassification as overload. In contrast, the discriminant scores of medium workload tend to decrease and drift down toward the separation line, with an increasing chance of classification as overload. In addition, the actual overload observations can be seen to

both increase and decrease as time passes. Specifically, the overload discriminant scores can be seen to drift down away from the separation line, and then appear to stabilize. Additionally, in some cases they start to drift back toward the separation line. As further evidence of this temporal effect Figure 4-1 has been rescaled in Figure 4-2 and shows only the first 15-minutes of scores. The temporal effect can be seen clearly in the middle and right side groups of scores. Specifically, the time trend in the low middle scores drifts up and away from the boundary, while the medium scores on the right drift down toward the "nominal vs. overload" boundary or separation line.

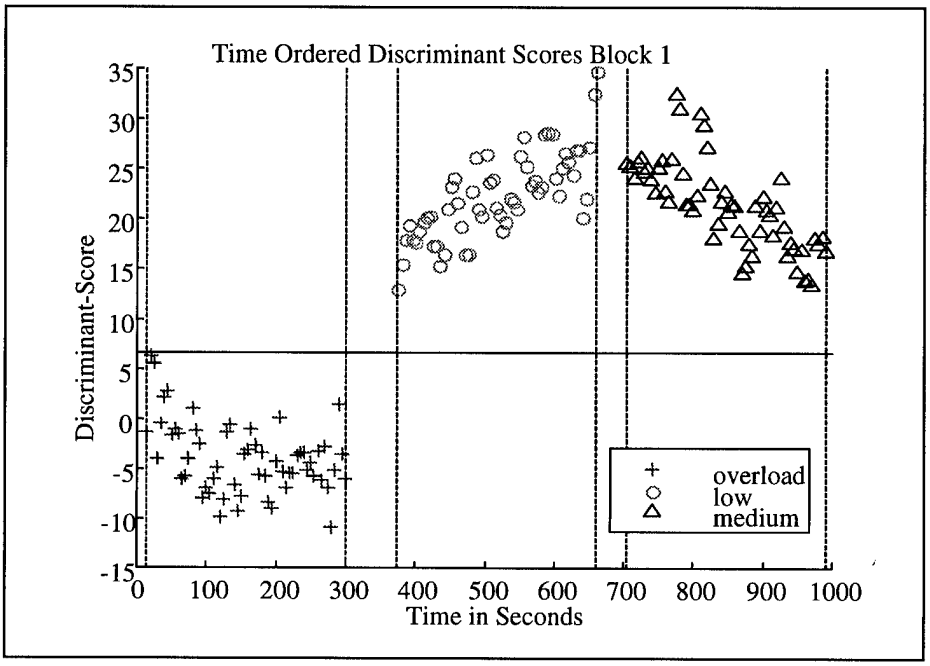

Figure 4-2. Rescaled Time Ordered Discriminant Scores.

4.1.2 Initial MLP ANN Model. A description of MLP ANNs including a discussion of architecture selection, training algorithms, and data set partitioning (training, test, and independent validation) was presented in Chapter 2. The initial MLP ANN modeling effort and all future ANN modeling were performed using *Matlab version 5.2* with the

Neural Network Toolbox version 3. All MLP ANNs in this research effort used a similar architecture with one input layer, one hidden layer and one output layer. As mentioned in Chapter 2, one input node is created for every input feature, and one for every output class. The only variable in network architecture was the number of hidden layer nodes. As a starting point, the number of hidden nodes was set equal to the number of input features. In addition, the hidden and output layer utilized a Log-Sigmoid activation function that is continuously differentiable and generates outputs between zero and one. Finally, as mentioned in Chapter 3, target vectors were set to 0.1 for false classes and 0.9 for the true class of an observation. Target values of 0.1 and 0.9 were selected over 0.0 and 1.0 to reduce training time. The initial model's architecture is summarized in the following table.

Table 4-1. Initial Network Architecture.

Layer	# of Nodes
Input	36
Hidden	36
Output	3

After the initial architecture was determined, parameters associated with training the network were determined. To assess feasibility for using SNR saliency screening, all weights were initialized close to zero using uniformly distributed values in a predefined range. A batch training algorithm that updated weights and biases after all exemplars were presented to the network was then selected. The algorithm incorporates momentum and an adaptive learning rate, as recommended by both Greene [19], Greene et. al. [20, 23] and Russell et. al. [38,39]. The momentum was initially set at 0.3 as would be used in SNR feature screening. Other parameters set included the maximum number of epochs

to train (which was initially set to 500) and the number of early stopping epochs. In the training algorithm, if the test set ANN error did not improve after a predetermined number of training epochs, the training was stopped and the weights and biases for the ANN with the smallest test set error were saved. The number of early stopping epochs was initially set to 20. Additionally, the data was normalized to have a mean of 0.0 and a standard deviation of 1.0. A summary of parameter setting follows as Table 4-2.

Table 4-2. Initial Parameter Settings.

Parameter	Setting
Range of Weight initialization	-0.001 to 0.001
Type of Learning Rate	Adaptive
Momentum	0.3
Type of Data Normalization	Gaussian
Early Stopping Epochs	20
Maximum Epochs	500

The final decision for training the ANN was the division of data between training, test, and validation sets. As mentioned in Chapter 3, the three blocks of low, medium, and overload observations appeared to provide a natural division of the data. Thus, as a first effort, the exemplars from the first block were used as the training set, the exemplars from the second set were used as the test set for early stopping, and the last set was used as an independent validation set. Figure 4-3 summarizes the training cycle by providing the sum of square error by epoch.

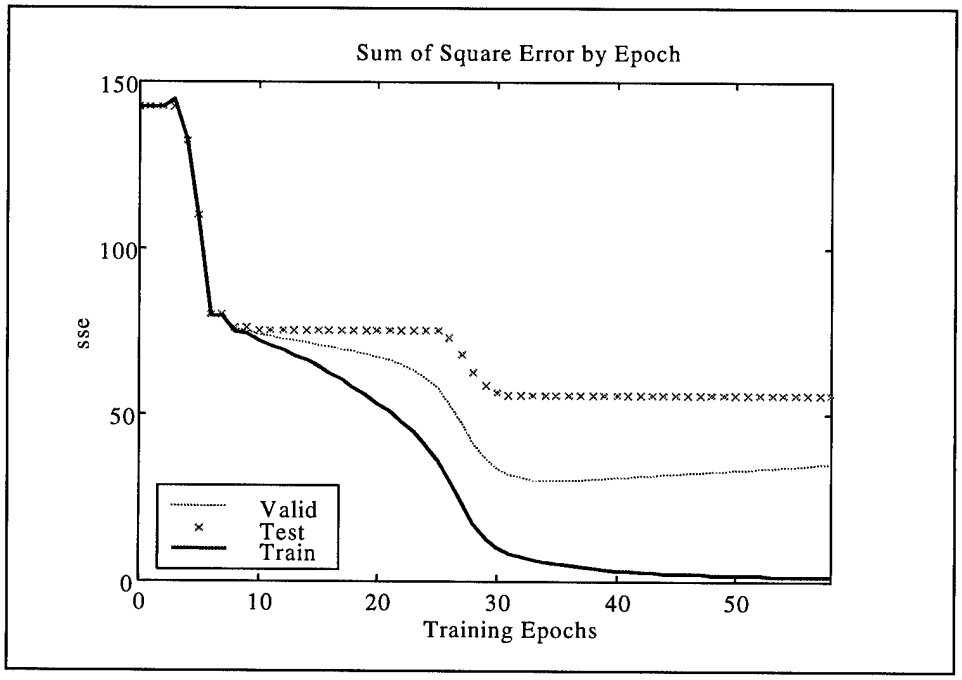

Figure 4-3. First ANN Training.

In this modeling effort, the training was stopped after 57 epochs. The optimal network was found after only 37 epochs, and corresponded to the network with the minimal error for the test set. From Figure 4-3, one abnormal training aspect was observed. As training proceeded, the validation set performed much better than the test set. While this is not necessarily unacceptable, it is suspicious. In order to utilize the trained ANN for classification, a "winner-take-all" approach was used in which the classification of an observation was assigned to the output node with the largest value. A classification accuracy (CA) for each model can then be defined as the percentage of correctly classified observations. For this particular model, the training set CA was 67%, the test set CA was 47% and the validation set CA was 61%. While all better than chance, none of these percentages are very impressive. Additional models were trained in which most showed a similar pattern of better performance by the validation set. Also,

some ANNs did not manage to successfully converge. These ANNs provided CA no better than chance in which each CA was 33% with all exemplars assigned to one class.

One hypothesis to explain the rather poor training of the ANNs is that each of the three blocks of observations may have its own unique pattern, possibly explained by an underlying pattern similar to that observed in the discriminant scores. To facilitate removal of this pattern, all three blocks of data were normalized separately. In doing so, the hope was that at least the mean values for each workload condition in each 15-minute block would be more consistent between blocks. In addition, data from the previously defined training and test sets were combined as one training/test set. Half of the data was then randomly selected to form the training set, with the remaining data assigned to the test set. The validation set remained as a separate 15-minute block of data. A plot of the network error for each of the three sets during training using the same parameters and architecture as the first ANN is presented as Figure 4-4.

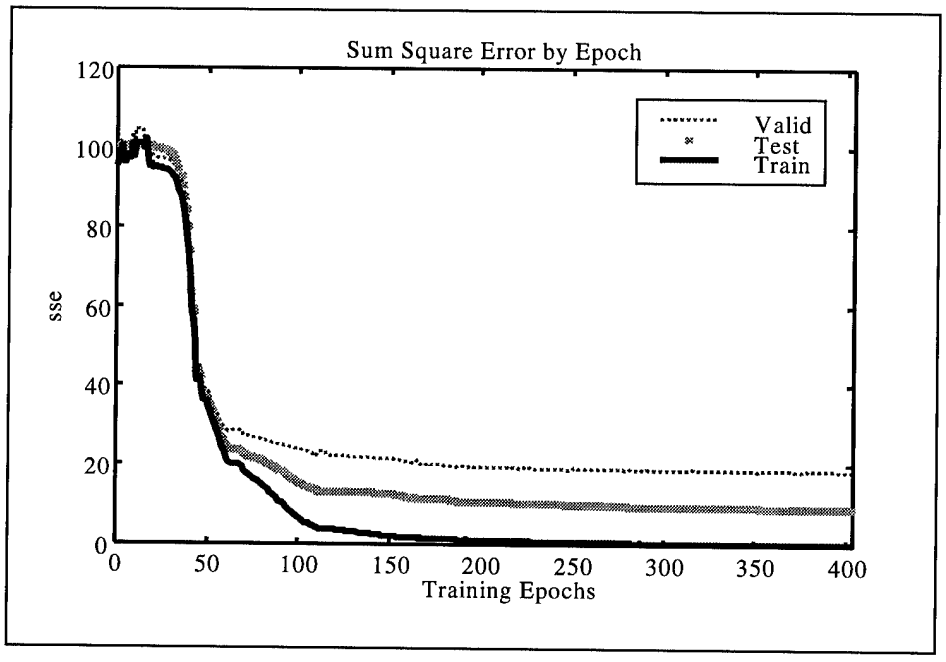

Figure 4-4. Second ANN Training Approach.

Figure 4-4 shows a much more typical pattern where the test set error is smaller than the validation set error. In addition, this ANN trained for nearly 400 epochs, while the first ANN's optimal weights were determined after only 37 epochs. Additionally, the error for all three sets is considerably smaller in this second effort. The resulting CA for each set also improved greatly. For Figure 4-4, the training set CA was 100%, the test set CA was 96%, and the validation set CA was 85%.

One possible explanation of the improved CA is that a homogeneous set of training and test exemplars is required for optimal training. If the training and test sets contain different patterns corresponding to the desired output, then the ANN may not be able to adequately train for classification outside of the training set. By including observations from two workload blocks a more generalized set of data was used for the training set. The training and training-test sets now include exemplars from a 30-minute vs. a 15-minute period. By expanding the training set and using a similarly constructed test set, the number of epochs increased before overfitting the training data. Therefore, for all future ANN training a homogenous training/test set composition will be utilized, i.e. the training and training-test sets will each include a random composition of exemplars from the two 15-minute blocks identified for model training for each subject. As eluded to, this will be done in the hope of finding an underlying pattern that will better fit the training-test set data, allow for a longer training period, and will hopefully be more applicable to the independent 15-minute block of validation data.

4.1.3 Consequences of Initial Efforts. As a result of the initial linear two-class discriminant modeling and the initial three-class ANN modeling, efforts will be made to

minimize the effects of temporal patterns and to ensure a homogenous set of training/test data is utilized. To minimize the temporal effects, each of the three 15-minute blocks of data will be normalized independently. In addition, all training and test sets will consist of randomly selected samples from a larger common set. Finally, to further minimize temporal effects, a systematic pseudo-Latin squares approach will be used to identify the validation set to be used by each subject. The CA of each validation set can then be used as the primary means to compare discriminant and ANN models. Assignment of the validation sets is shown in the Table 4-3.

Table 4-3. Validation Set Assignment.

Subject	Block 1	Block 2	Block 3
02	Train	Validation	Train
03	Train	Train	Validation
04	Validation	Train	Train
05	Train	Validation	Train
09	Train	Train	Validation
11	Validation	Train	Train
13	Train	Validation	Train
16	Train	Train	Validation

4.2 Individual Discriminant Models.

Discriminant modeling will be used as a benchmark to compare the performance of ANN classification and as a means to select salient features for ANN use. This section provides the methodology used for the three-class discriminant model. Two heuristic methods for feature reduction are also presented. A summary of the results obtained from the models are then included.

As identified in Table 4-3, two blocks of data can be used to "train" the discriminant models. The validation set will be used to assess the classification accuracy.

The training data will essentially be used to estimate the first two statistical moments for each of the three workload populations. Here the vector of mean values for each input feature is the 1st moment and the covariance matrix of all input features is the 2nd moment. Equation 2-11 will then be used to compute a d^Q score for each of the three classes for each observation. The observation is then assigned to the class with the largest d^Q score that corresponds to the greatest probability of class membership.

4.2.1 Feature Selection by Coefficient. While determining the parameters in Equation 2-11 is relatively straight forward, determining the optimal set of input features is not. For this research effort, any combination of the 36 psychophysiological may be used. As mentioned in Chapter 2, the input feature coefficients from the discriminant function can be used as one measure of saliency if all input features are normalized. The resulting coefficients are unitless and their magnitude provides a measure of their saliency in the model. A derivation of Equation 2-11 is presented by Bauer [3] to determine coefficients as follows.

Starting with Equation 2-11:

$$d_k^Q(\mathbf{x}) = -\ln|\Sigma_k| - (\mathbf{x} - \mu_k)'\Sigma_k^{-1}(\mathbf{x} - \mu_k) + \ln P_k$$

define a constant, $c_k = -\ln|\Sigma_k| + \ln P_k$, for each group,

$$d_k^Q(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_k)'\Sigma_k^{-1}(\mathbf{x} - \mu_k) + c_k \qquad (4\text{-}4)$$

multiply terms,

$$d_k^Q(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}'\Sigma_k^{-1}\mathbf{x} - \mu_k'\Sigma_k^{-1}\mathbf{x} - \mathbf{x}'\Sigma_k^{-1}\mu_k + \mu_k'\Sigma_k^{-1}\mu_k) + c_k \quad (4\text{-}5)$$

collect terms,

$$d_k^Q(\mathbf{x}) = -\frac{1}{2}(\mu_k'\Sigma_k^{-1}\mu_k) - \frac{1}{2}(\mathbf{x}'\Sigma_k^{-1}\mathbf{x} - 2\mathbf{x}'\Sigma_k^{-1}\mu_k) + c_k \quad (4\text{-}6)$$

assume $\Sigma_k = \Sigma \; \forall \; k$, $\mathbf{x}'\Sigma_k^{-1}\mathbf{x}$ is constant across all groups (the quadratic term is equivalent for all groups). Therefore, revising the group constant $c_k' = c_k - \frac{1}{2}(\mathbf{x}'\Sigma^{-1}\mathbf{x})$,

$$d_k^Q(\mathbf{x}) = -\frac{1}{2}(\mu_k'\Sigma^{-1}\mu_k) - \mathbf{x}'\Sigma^{-1}\mu_k + c_k' \quad (4\text{-}7)$$

Note, $\mu_k'\Sigma^{-1}\mu_k$ is a constant for each group, revise the group constant a last time as $c_k'' = c_k' - \frac{1}{2}(\mu_k'\Sigma^{-1}\mu_k)$ then,

$$d_k^Q(\mathbf{x}) = c_k'' - \mathbf{x}'\Sigma^{-1}\mu_k \quad (4\text{-}8)$$

Finally, let $b_k = c_k''$ and $\mathbf{b}_k = \Sigma^{-1}\mu_k$. A more familiar form of the discriminant function is derived where b_k is some constant value for each group, and $\mathbf{b}_k$ is the vector of coefficients for input variables.

$$d_k^Q(\mathbf{x}) = c_k'' - \mathbf{x}'\Sigma_k^{-1}\mu_k = b_k - \mathbf{x}'\mathbf{b}_k \quad (4\text{-}9)$$

Thus for k populations, k vectors of coefficients are obtained.

The coefficients obtained from the normalized input features can now be used for saliency screening. A methodology for ordering the features by linear saliency is as follows:

1. Using all input features, "train" the discriminant functions and compute a vector of coefficients for each of k populations.
2. From the set of k coefficient vectors find the maximum absolute coefficient value for each input feature.
3. Remove the input feature with the minimum of the maximum absolute values.
4. Using the reduced set of features, "re-train" the discriminant functions and repeat steps 2 and 3 until all features have been removed.

In step 2, the maximum absolute value for each input variable's coefficient was found. This value was selected as a quantitative means to assess an input feature's strongest predictive power. Thus, if an input feature had a large associated coefficient for one of the k groups, it was considered to be salient. In step 3, the feature with the minimum of all the maximum absolute values was then removed. Overall, the intuition here was to formulate an initial model using all input features, then to remove features one-by-one in a fashion that would select the least salient feature to remove during each iteration. Additionally, by starting with all input features, all input variable interactions were initially included. Then, each time a variable was removed, the remaining subset would include not only the best set of individually salient variables, but also an optimal composition of input variables with one less variable than the previous input set.

To determine an optimal parsimonious set of salient features the classification accuracy for each discriminant model with a reduced set of features can be used as a measure of effectiveness (MOE). This MOE can then be reviewed to see where the first significant reduction in training set CA occurs. The feature that caused significant reduction in CA along with all features removed after it should then be retained as salient.

In addition, a feature of noise can also be added to the set of input features to determine which features appear to be more salient than noise.

A plot of classification accuracy for the three-class discriminant model for subject 09 with feature reduction by coefficient saliency is shown in Figure 4-5.

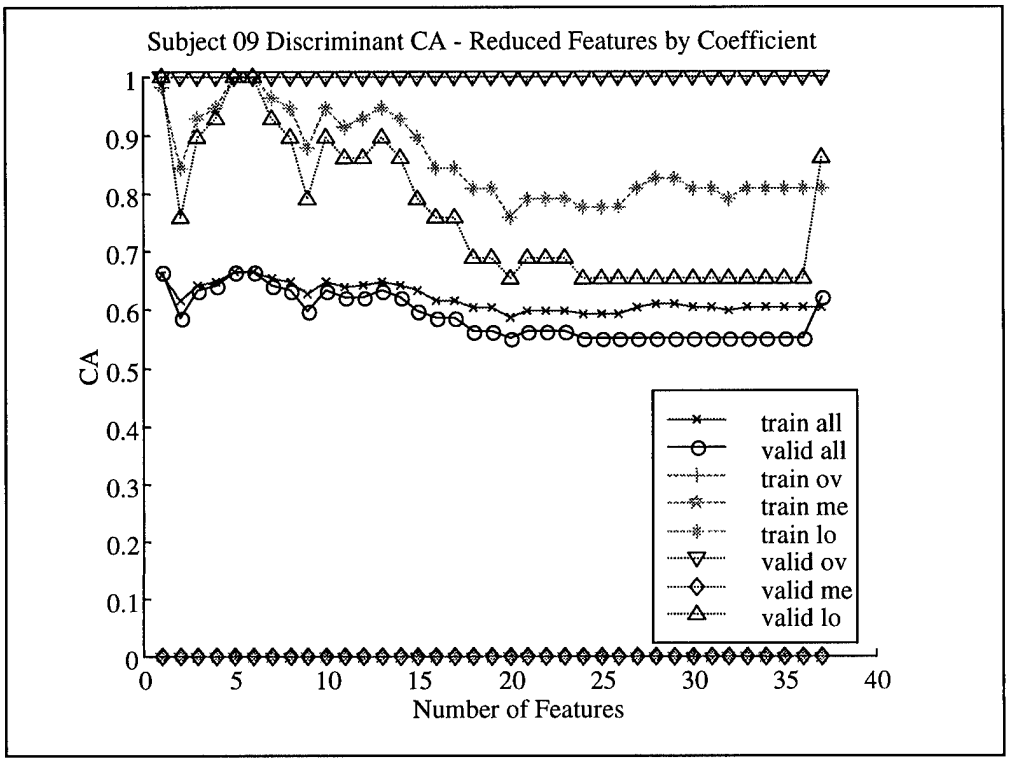

Figure 4-5. Feature Reduction by Coefficient.

Figure 4-5 shows a classification accuracy for both the training and validation sets for low and overload conditions to be 100% when using 5 or 6 features. Although difficult to see, the training set overload CA as denoted by "+'s" are directly under the validation set triangular symbols and all remain at 100% CA as features are reduced. Unfortunately, the overall CA was only 66% at this point due to misclassification of every medium observation. In general, for all eight subjects, a similar plot was obtained

where low and overload classification was above 90%, while medium classification remained close to 0%.

4.2.2 Feature Selection by Loading. As a second measure of linear saliency, a similar screening method was employed that uses the maximum discriminant loading magnitude rather than the maximum coefficient magnitude for each variable. After determining the coefficients for each input variable, the discriminant loading can be computed using Equation 2-12. A methodology for ordering the features by linear saliency using loadings is as follows:

1. Using all input features, "train" the discriminant functions and compute a vector of coefficients for each of k populations.
2. Calculate k vectors of discriminate loadings.
3. From the set of k loadings, find the maximum absolute loading for each input feature.
4. Remove the input feature with the minimum of the maximum absolute values.
5. Using the reduced set of features, "re-train" the discriminant functions and repeat steps 2, 3, and 4 until all features have been removed.

Logic similar to that used for the discriminant function coefficient saliency screening heuristic is used to explain the intuition behind this saliency screening heuristic as well. The only difference is the use of the discriminant loading as the measure of a feature's saliency as opposed to using the magnitude of the coefficient.

As in coefficient screening, an optimal parsimonious set of salient features can be determined for the discriminant model by using CA. The CA can be reviewed to identify the first significant reduction in the training set CA. The feature that caused significant reduction in CA along with all features removed after it should then be retained as salient. In addition, a feature of noise can be injected to determine which features appear to be

more salient than noise. A plot of CA for the same subject with feature reduction by discriminant loadings is shown in Figure 4-6.

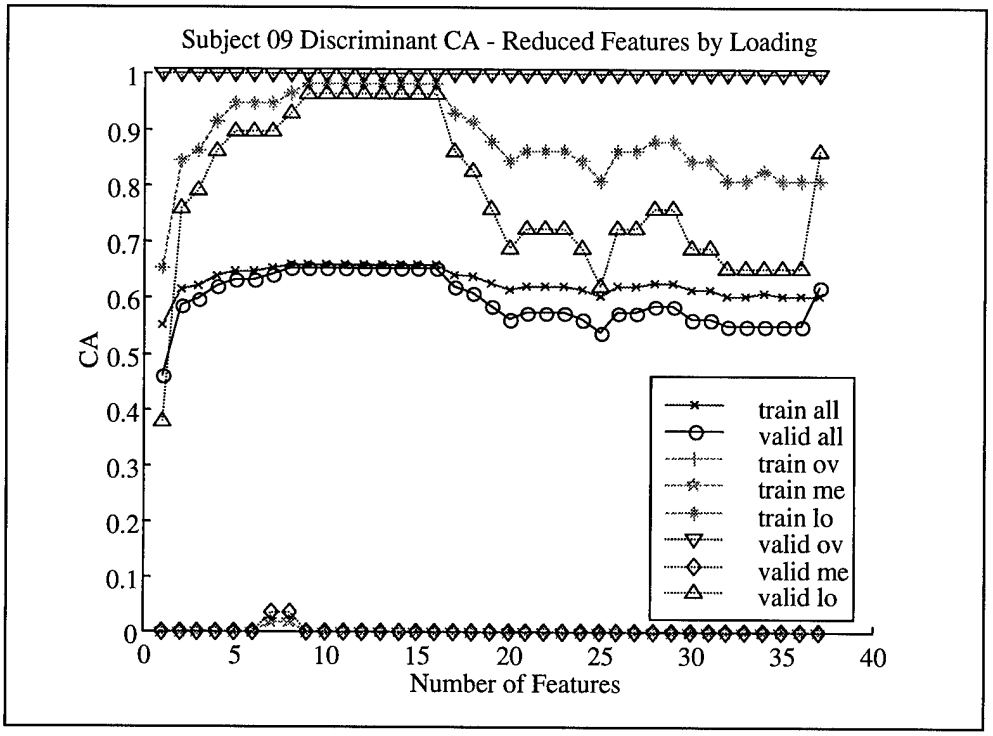

Figure 4-6. Feature Reduction by Loadings.

As was seen in Figure 4-5, the overall classification for training and validation set low and overload observations is close to 100% for a range of input features used. Again, the training set overload scores are difficult to see, but remain at 100% throughout the feature reduction. Unfortunately, once again, almost all medium observations were misclassified. The most notable difference between the two plots is the number of optimal input features to retain. When removing features by coefficient saliency, 5 or 6 features appeared to be optimal. When removing features by discriminant loadings, 8 to 10 features appear optimal. To compare salient features obtained by both methods, the

Table 4-4 shows the top 10 features, rank ordered by both loading and coefficient saliency.

Table 4-4. Feature Rank by Linear Saliency.

Linear Saliency		
Rank	*Loading*	*Coefficient*
1	**F7-ub**	**PZ-ub**
2	**PZ-ub**	**F7-ub**
3	**F7-b**	**PZ-b**
4	FZ-ub	**O2-ub**
5	T8-ub	**F7-b**
6	**O2-ub**	**IBRI**
7	T7-ub	T8-ub
8	IBLI	T8-b
9	F7-a	T7-ub
10	**IBRI**	FZ-ub

A quick glance at Table 4-4 shows that both saliency screening methods are consistent. The top two features are identical and eight of the top ten features are identical in each case. In addition, the top six coefficient screening features are highlighted in bold print. Five of these six features are then highlighted in the list of features selected by discriminant loadings, with F7-b not present among the features chosen by loading. In addition, because discriminant loadings are independent of input variable correlations two features that would have the same relative predictive power will be ranked together when using loadings as the saliency measure, while they may not be when using coefficients. This is true because if multicolinearity and high correlations are present, an otherwise salient feature may have inconsistent or smaller than expected coefficient values. A smaller set of optimal features would then be expected when screening by coefficients. Thus, the top ten features determined by loading may have the

same predictive power as the top six features determined by coefficients when used in a discriminant model. On-the-other-hand, when screening for salient features for use in an MLP ANN or another more complex classification model, the larger set of linearly salient features identified by loadings may be preferred.

4.2.3 Summary of Discriminant Analysis. After computing discriminant models with input feature saliency screening based on coefficients and loadings for all eight test subjects, similar patterns repeatedly occurred. First, none of the models classified medium workload well. After looking at the mean values of subject 09's input features and the diagonal of the covariance matrix, the cause of poor medium classification appeared to be the mean values of the medium population. Overall, the variance of the medium features did not appear significantly different from low or overload. Additionally, all medium mean values were between the low and overload means, but depending on the input variable the mean was either very close to the low, or close to the overload value. Thus, when classifying a medium observation, depending on the features used, it was likely to be classified as low or overload, especially if the medium covariance structure was slightly larger than the low or overload covariance structure. The second pattern derived from the discriminant analysis involved the salient features identified. Not only did individuals identify a consistent set of features using both loading and coefficient screening, but a pattern of top features occurred between all subjects. This set of "universally" salient linear features was dominated by ultrabeta features and eye-blink features (IBLI and blnks) followed by beta features and IBRI. All ultrabeta and eye-blink features ranked in the top 10 are shown in Table 4-5 and are highlighted in bold print.

Table 4-5. Salient Linear Features by Subject.

Rank	Subj 02		Subj 03		Subj 04		Subj 05		Subj 09		Subj 11		Subj 13		Subj 16	
	Load	Coeff	Load	Coeff	Load	Coeff	Load	Coeff	Load	Coeff	Load	Coeff	Load	Coeff	Load	Coeff
1	**blnks**	**IBLI**	**PZ-ub**	**PZ-ub**	**O2-ub**	**O2-ub**	IBRI	IBRI	**F7-ub**	**PZ-ub**	T8-b	T8-b	**blnks**	**IBLI**	**T7-ub**	**T7-ub**
2	**IBLI**	**blnks**	**FZ-ub**	**FZ-ub**	O2-b	**PZ-ub**	**T7-ub**	**T7-ub**	**PZ-ub**	**F7-ub**	**T8-ub**	**T8-ub**	**IBLI**	**blnks**	T7-b	**O2-ub**
3	**FZ-ub**	**FZ-ub**	**O2-ub**	IBRI	**IBLI**	**IBLI**	T7-b	**O2-ub**	F7-b	PZ-b	**PZ-ub**	**F7-ub**	**O2-ub**	T8-a	**O2-ub**	T7-b
4	HR	**T8-ub**	**T7-ub**	**T7-ub**	T8-b	T8-b	**FZ-ub**	**PZ-ub**	**FZ-ub**	**O2-ub**	**O2-ub**	**FZ-ub**	T8-a	**O2-ub**	**PZ-ub**	**T8-ub**
5	**O2-ub**	O2-a	**T8-ub**	T7-b	**T8-ub**	T7-t	**O2-ub**	**F7-ub**	**T8-ub**	F7-b	T7-b	**O2-ub**	T7-a	T8-b	**F7-ub**	FZ-b
6	**T8-ub**	**T7-ub**	O2-t	**T8-ub**	**F7-ub**	**T8-ub**	**PZ-ub**	brths	**O2-ub**	IBRI	**FZ-ub**	FZ-b	F7-a	**F7-ub**	**FZ-ub**	FZ-t
7	T8-b	**F7-ub**	**F7-ub**	F7-b	T7-b	**T7-ub**	**T8-ub**	T7-b	**T7-ub**	**T8-ub**	O2-b	O2-b	HrVar	**T7-ub**	F7-b	F7-b
8	**T7-ub**	FZ-t	PZ-t	F7-d	**T7-ub**	T7-b	**IBLI**	PZ-b	**IBLI**	T8-b	**T7-ub**	F7-b	O2-b	FZ-t	FZ-b	T7-t
9	**PZ-ub**	T8-b	PZ-b	O2-d	**PZ-ub**	IBRI	T8-b	**FZ-ub**	F7-a	**T7-ub**	PZ-b	F7-t	O2-t	F7-a	HR	F7-d
10	PZ-b	O2-b	T8-b	brths	**blnks**	T8-t	brths	**T8-ub**	IBRI	**FZ-ub**	FZ-b	O2-a	FZ-b	FZ-d	O2-b	T7-d
11	FZ-t	PZ-t	FZ-b	T8-b	PZ-b	FZ-t	F7-ub	O2-a	PZ-b	FZ-b	FZ-t	T7-ub	T7-b	F7-d	FZ-t	F7-a
12	T7-b	O2-d	FZ-t	O2-ub	FZ-ub	O2-d	O2-t	F7-t	F7-t	O2-b	F7-t	FZ-a	PZ-b	T7-d	F7-t	O2-t
13	O2-b	PZ-d	T7-b	T8-a	HR	O2-b	blnks	HrVar	blnks	PZ-t	F7-a	IBLI	T8-b	PZ-d	FZ-a	PZ-t
14	T8-t	PZ-b	T8-t	FZ-d	O2-a	T8-d	PZ-t	T7-d	brths	O2-t	PZ-t	**noise**	T7-d	brths	T8-ub	PZ-ub
15	F7-b	T8-a	F7-b	F7-ub	HrVar	brths	FZ-t	O2-b	T7-t	T7-a	FZ-a	PZ-t	PZ-a	IBRI	IBLI	FZ-ub
16	O2-a	T7-d	O2-b	T7-d	FZ-t	PZ-b	HR	FZ-t	T8-t	F7-a	F7-ub	T7-d	PZ-ub	T7-b	PZ-b	T8-b
17	F7-ub	F7-d	T8-a	F7-t	T8-t	blnks	O2-b	T7-t	T8-b	blnks	O2-t	F7-d	F7-b	PZ-b	brths	PZ-a
18	F7-d	O2-t	PZ-a	T8-t	T8-a	O2-a	T7-t	F7-a	FZ-b	IBLI	blnks	PZ-a	T8-t	O2-b	blnks	T8-t
19	O2-t	**noise**	FZ-d	T7-a	brths	F7-ub	F7-a	T7-a	HR	HR	HR	O2-d	PZ-t	PZ-ub	T7-a	HrVar
20	F7-a	T8-d	F7-a	FZ-b	F7-b	FZ-b	F7-t	HR	T7-a	T8-a	F7-b	F7-a	O2-a	F7-t	T8-t	F7-ub
21	F7-t	HR	FZ-a	T7-t	F7-t	T7-d	F7-d	T8-d	FZ-t	F7-t	IBLI	FZ-d	T7-t	PZ-t	PZ-t	IBRI
22	FZ-d	FZ-a	T7-a	**noise**	T8-d	F7-d	FZ-a	PZ-a	O2-t	T7-t	T7-t	PZ-d	T8-d	O2-a	T8-b	brths
23	T7-t	PZ-ub	IBRI	HrVar	PZ-t	**noise**	T7-a	FZ-a	PZ-a	FZ-t	FZ-d	IBRI	FZ-t	HR	T8-a	PZ-b
24	PZ-t	T8-t	T7-t	IBLI	O2-t	FZ-ub	O2-a	IBLI	PZ-d	brths	O2-a	brths	F7-d	T7-t	FZ-d	T7-a
25	T8-a	HrVar	F7-t	PZ-d	IBRI	FZ-a	FZ-b	T8-b	T7-d	F7-d	O2-d	HR	F7-t	T8-t	T8-d	IBLI
26	PZ-a	brths	O2-d	PZ-a	T7-t	F7-t	T7-d	O2-d	F7-d	T7-d	PZ-d	T7-t	brths	T8-ub	O2-a	O2-a
27	T7-a	F7-b	O2-a	O2-a	FZ-a	HR	F7-b	PZ-d	PZ-t	PZ-d	T8-a	T7-b	IBRI	FZ-ub	F7-d	FZ-a
28	T7-d	F7-t	brths	FZ-t	F7-a	T8-a	T8-d	T8-a	FZ-d	FZ-d	PZ-a	PZ-ub	FZ-a	O2-d	F7-a	HR
29	O2-d	T7-t	blnks	PZ-t	T7-a	PZ-d	PZ-b	O2-t	O2-d	T8-d	IBRI	O2-t	FZ-ub	F7-b	PZ-a	O2-d
30	FZ-b	T7-b	HrVar	PZ-b	F7-d	T7-a	T8-a	FZ-d	T8-a	HrVar	T8-d	FZ-t	O2-d	FZ-a	O2-t	PZ-d
31	IBRI	PZ-a	PZ-d	blnks	PZ-d	HrVar	O2-d	PZ-t	T8-d	T7-b	T7-d	T8-a	F7-ub	HrVar	PZ-d	FZ-d
32	FZ-a	FZ-b	IBLI	O2-t	T7-d	FZ-d	PZ-a	blnks	O2-b	PZ-a	T7-a	PZ-b	FZ-d	T8-d	IBRI	**noise**
33	T8-d	O2-ub	**noise**	FZ-a	FZ-d	PZ-t	T8-t	T8-t	**noise**	O2-a	T8-t	T8-d	PZ-d	FZ-b	T7-d	O2-b
34	HrVar	IBRI	T8-d	F7-a	PZ-a	F7-a	PZ-d	F7-b	O2-a	T8-t	F7-d	T7-a	T7-ub	O2-t	O2-d	T8-d
35	**noise**	F7-a	F7-d	T8-d	FZ-b	PZ-a	FZ-d	**noise**	HrVar	**noise**	**noise**	T8-t	HR	PZ-a	T7-t	F7-t
36	PZ-d	T7-a	T7-d	O2-b	**noise**	F7-b	HrVar	FZ-b	T7-b	FZ-a	brths	blnks	T8-ub	T7-a	HrVar	T8-a
37	brths	FZ-d	HR	HR	O2-d	O2-t	**noise**	F7-d	FZ-a	O2-d	HrVar	HrVar	**noise**	**noise**	**noise**	blnks

In summary, discriminant analysis classification could not detect medium workload well which may suggest the need of a more complex model such as an MLP ANN. Additionally, as a starting point, those features identified as linearly salient may be used as a foundation set to build upon for use in the MLP ANNs.

4.3 Individual ANN Models

Unlike linear discriminant models, MLP ANNs do not produce the same results every time they are trained. For this reason, an optimal set of salient features must be

selected, then the ANN must be trained multiple times to obtain a mean value for the model's ability to classify.

4.3.1 SNR Saliency Screening. As presented in Chapter 2, the SNR saliency measure can be used to screen salient features while training an ANN. While the methodology presented is Sumrell's [45], a slight change was made in step 3 for this research effort. The revised steps are as follow:

1. Add a noise feature, x_N, to the original set of features.
2. Begin training of the neural network.
3. Interrupt training after the salieny metric values have stabilized. Assume the saliency metric has stabilized if network is trained.
4. Identify the feature with the lowest SNR value and remove it from further training.
5. Continue training the neural network.
6. Repeat steps 3, 4, and 5 until all of the features in the original set have been removed.
7. Compare the reaction of the test set classification error rate to the removal of the individual features. Retain the first feature whose removal caused a significant increase in the test set classification error rate, as well as all features that were removed after that first salient feature.

The only significant change is to assume that the saliency measure has stabilized after the network has been trained to a set number of epochs, or until test set performance does not improve for a set number of epochs. While completely training the ANN after each feature is removed may result in an increased number of training epochs, the computational operations required to compute the SNR measure and to assess when this measure has stabilized may make training times slower, even though fewer epochs are included in the screening process. In addition, by setting the test set stopping epochs as small as possible, training may end at local minima were the least salient feature is removed before training resumes. Using normalized input features by block, and a homogenous training and test set, SNR screening was performed for all subjects at least

three times with 18, 36, and 72 hidden nodes. A summary of the MLP ANN architecture used is defined in the Table 4-6 below.

Table 4-6. SNR ANN Architecture.

Layer	# of Nodes
Input Nodes	37 to 2
Hidden Nodes	18, 36, and 72
Output Nodes	3

Other ANN parameters were set as identified in Table 4-2. Figure 4-7 shows the network performance by epoch using subject 09 with 36 hidden nodes as features are removed.

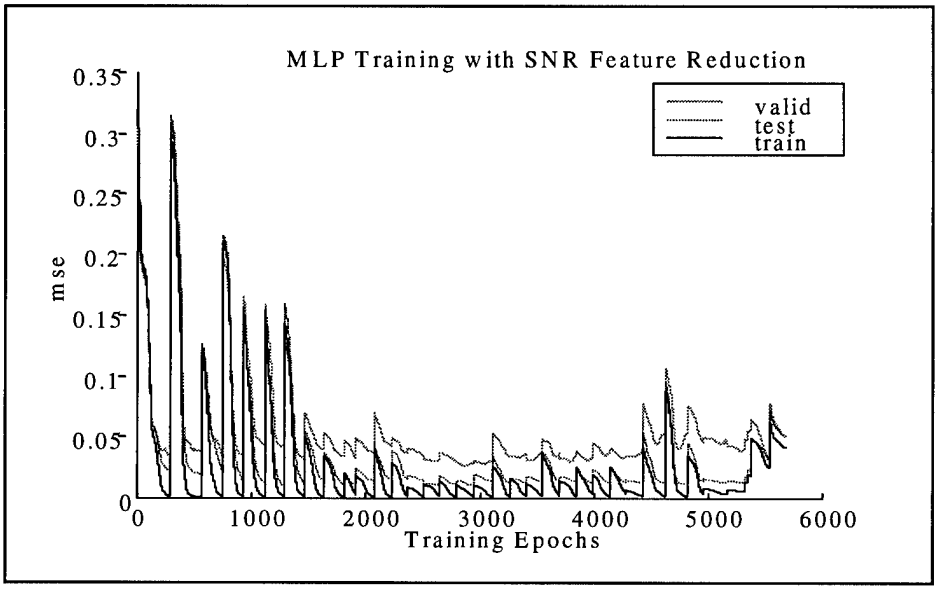

Figure 4-7. SNR Feature Screening.

From Figure 4-7, each spike in mean square error (MSE) corresponds to the removal of a feature from the network. After a feature is removed, the network quickly trains to achieve an optimal MSE, if enough features are still in the model. The effect of not having enough features can be seen somewhere around 5000 training epochs, after a

majority of the features have been removed. To assess the number of features to retain, plots of classification accuracy obtained during SNR feature reduction can be analyzed to determine the point of significant classification degradation. The plot of CA corresponding to the above training session is presented as Figure 4-8. Here approximately eight input features appear to provide the salient set as seen by the drop in CA (reading right to left) as features are removed from the model.

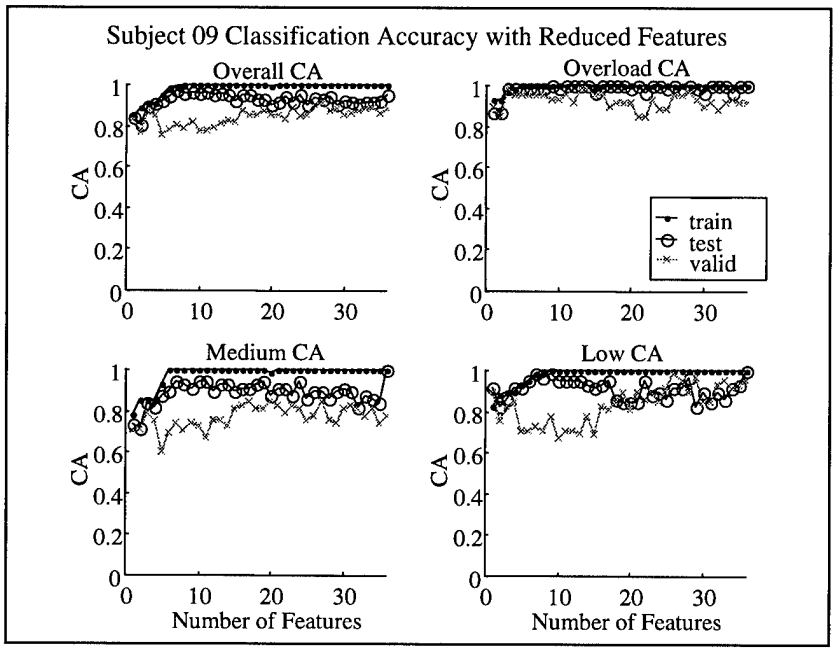

Figure 4-8. CA with SNR Feature Reduction.

After performing the three SNR screening runs, these input feature rank orderings can be added to the two feature orderings derived from the linear screening. A total of five separately ordered salient features are now available to select an optimal parsimonious set of salient features for each individual. As was suggested in performing linear feature selection, the features should be retained prior to a significant decrease in training set CA. While this could be performed in one run for the linear models, each

ANN performed slightly differently depending on the training/test set composition. Additionally, weights for the ANN were randomly initialized close to zero, which could also impact the results with a non-deterministic gradient descent search for the minimum error. Thus, selecting an absolute optimal set of features for each individual is not the goal. The task at hand is simply to select an individually tailored set of features that appeared promising for predicting workload in the independent validation set. Table 4-7 shows the top 15 features suggested by all five models for subject 09. The optimal set of features selected contains eight features as highlighted in bold print.

Table 4-7. Subject 09 Top 15 Features.

Rank	*Linear Loading*	*Linear Coefficient*	*SNR ANN 18 Nodes*	*SNR ANN 36 Nodes*	*SNR ANN 72 Nodes*
1	**F7-ub**	**PZ-ub**	**PZ-ub**	**PZ-ub**	**PZ-ub**
2	**PZ-ub**	**F7-ub**	**F7-ub**	**FZ-ub**	**F7-ub**
3	F7-b	**PZ-b**	**O2-ub**	**F7-ub**	**O2-ub**
4	**FZ-ub**	**O2-ub**	**T7-ub**	F7-b	PZ-b
5	**T8-ub**	F7-b	**T8-ub**	**O2-ub**	**T8-ub**
6	**O2-ub**	IBRI	**FZ-ub**	HR	O2-d
7	**T7-ub**	**T8-ub**	**IBLI**	**PZ-b**	FZ-a
8	**IBLI**	T8-b	**PZ-b**	T8-t	O2-b
9	F7-a	**T7-ub**	HR	**T7-ub**	**IBLI**
10	IBRI	**FZ-ub**	IBRI	PZ-d	**T7-ub**
11	**PZ-b**	FZ-b	T8-a	O2-d	HR
12	F7-t	O2-b	O2-b	IBRI	F7-t
13	blnks	PZ-t	F7-b	T8-b	T7-a
14	brths	O2-t	T8-d	brths	T8-t
15	T7-t	T7-a	PZ-a	HrVar	**FZ-ub**

As previously mentioned, this set is not guaranteed to be the best possible set, but should perform well. After performing at least three SNR ANN screenings of all eight individuals, similar feature ranking results as those presented for subject 09 occurred.

Specifically, most feature ranks for an individual contained the same top feature. Additionally, if the top feature was not identical, then the top two features were. This was consistent across linear models and non-linear ANNs, and across the architecture of the number of hidden nodes selected. In addition, most models showed a threshold in the training set CA performance with between 5 and 15 features remaining in the model. Finally, as shown in Table 4-7, many of the top 15 features were consistent to all five feature rankings. A summary of the heuristic method of selecting the top features to be used for each individual is as follows:

1. Perform linear screening using both coefficients and loadings as saliency measures.
2. Perform 3 SNR ANN feature saliency training runs.
3. Determine minimum number of features before significant test set CA degragation.
4. Compare CA identified in step 3 for the three SNR screening runs.
5. Compare those features utilized by the model with the most "robust" CA in step 4 to those obtained by the other four models.
6. Use the features as selected by the most "robust" model if consistent with other models or augment with additional features consistently rated highly by other models.

In the specific case of subject 09, the SNR ANN with 18 hidden nodes appeared to have the most robust CA with a reduced set of features. As used in this feature selection methodology, robust refers to a CA that is not only large in magnitude, but also appears consistent and hopefully reproducible. For example, if the best CA for one model was found using 7 features, but very poor performance was obtained when using 6 or 8 features, a separate model with a slightly worse CA may be preferred. Even with a lower CA, the second model is favored with demonstrated robustness over a range of more features. A summary of the selected salient set of parsimonious features by subject is given in Table 4-8.

Table 4-8. Salient Features by Individual.

Feature	Name	Individually Selected Salient Features s02	s03	s04	s05	s09	s11	s13	s16	Times Selected
1	O2-d									0
2	O2-t		X							1
3	O2-a	X	X							2
4	O2-b			X						1
5	O2-ub		X	X		X		X	X	5
6	PZ-d									0
7	PZ-t									0
8	PZ-a									0
9	PZ-b					X				1
10	PZ-ub		X			X	X		X	4
11	T7-d			X						1
12	T7-t			X						1
13	T7-a		X					X		2
14	T7-b	X		X	X				X	4
15	T7-ub		X		X	X				3
16	T8-d									0
17	T8-t									0
18	T8-a	X	X					X		3
19	T8-b			X	X		X			3
20	T8-ub		X	X		X	X	X	X	6
21	F7-d									0
22	F7-t									0
23	F7-a								X	1
24	F7-b		X						X	2
25	F7-ub	X		X	X	X	X		X	6
26	FZ-d									0
27	FZ-t	X			X			X	X	4
28	FZ-a		X							1
29	FZ-b	X		X						2
30	FZ-ub	X	X			X			X	4
31	HR	X	X							2
32	HrVar									0
33	blnks	X						X		2
34	IBLI	X		X		X	X	X		5
35	brths									0
36	IBRI		X	X						2
	Total Features:	10	13	11	5	8	5	7	9	

4.3.2 ANN Training with Optimal Features. Once the set of optimal features was selected, one ANN was trained multiple times to provide an estimate of the network's

true performance. In order to estimate the mean CA and associated standard deviation for each ANN, each model was trained at least 30 times. With randomly selected exemplars creating the training and test sets, each of the 30 runs provided CA values that can be assumed to be independent and identically distributed samples from a distribution with a population mean μ and finite standard deviation σ. Then, according to the Central Limit Theorem (CLT), with a sufficient number of samples (usually $n > 30$), the distribution will converge to a standard normal (Gaussian) distribution [47]. In addition, once the mean and associated standard deviation are determined, a confidence interval (CI) for the mean can also be computed. A formula for computing the CI of the mean is as follows:

$$100(1 - \alpha)\% \text{ CI} = \mu \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{4-10}$$

where α determines the level of confidence, μ is the estimated distribution mean, σ is the estimated standard error, $z_{\alpha/2}$ is the associated value of a standard normal distribution, and n is the number of samples. For 30 runs a 95% CI can be found as follows:

$$95\% \text{ CI} = \mu \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = \mu \pm 1.96 \frac{\sigma}{\sqrt{30}} \tag{4-11}$$

The above CI simply defines the lower and upper bounds in which the actual mean of the distribution should be found with 95% confidence.

When performing multiple runs of a single ANN to develop CIs for the different CAs, a standard architecture was used that set the number of hidden nodes to be twice the number of input features. For subject 09, since 8 input features were being used, the hidden layer was initialized to have 16 hidden nodes. Also, because the SNR ratio of all variables was not being computed, a more efficient algorithm was used to initialize the

network weights. This algorithm was developed by Nguyen and Widrow and assumes each hidden node is responsible for approximating a small portion of the desired output. Thus, all hidden nodes are initialized within the active region of the transfer functions to provide a piece-wise linear approximation to the desired targets. In practice, use of this initialization algorithm can reduce training time by an order of magnitude [13]. Also, to facilitate faster learning, the momentum rate was increased from 0.3 to 0.9. Finally, the maximum training epochs and the test set epochs were both increased from the SNR parameter settings to provide the ANN an optimal environment to achieve maximum performance on the test set. Table 4-9 summarizes the parameter settings used for each individually trained network used to calculate CA CIs.

Table 4-9. Individual ANN Parameters.

Parameter	Setting
Weight initialization	Nguyen-Widrow
Type of Learning Rate	Adaptive
Momentum	0.9
Type of Data Normalization	Gaussian
Early Stopping Epochs	50
Maximum Epochs	1000

A slight change was also performed while dividing the 30-minute set of training data. As before, the two blocks of data were randomly permuted with the first half comprising the training set. Next, half of the remaining data was used as the test set for early stopping. The remaining 25% of the data was then labeled as an internal validation data set to provide a measure the CA obtained from samples that are homogenous to those used for training and test. As identified in table 4-3, the last 15-minute block of data was held out from the training and test sets to be used as an external validation set.

Results of training an ANN 30 times using random permutations of the 30-minute training, test, and internal validation data are presented in Figure 4-9.

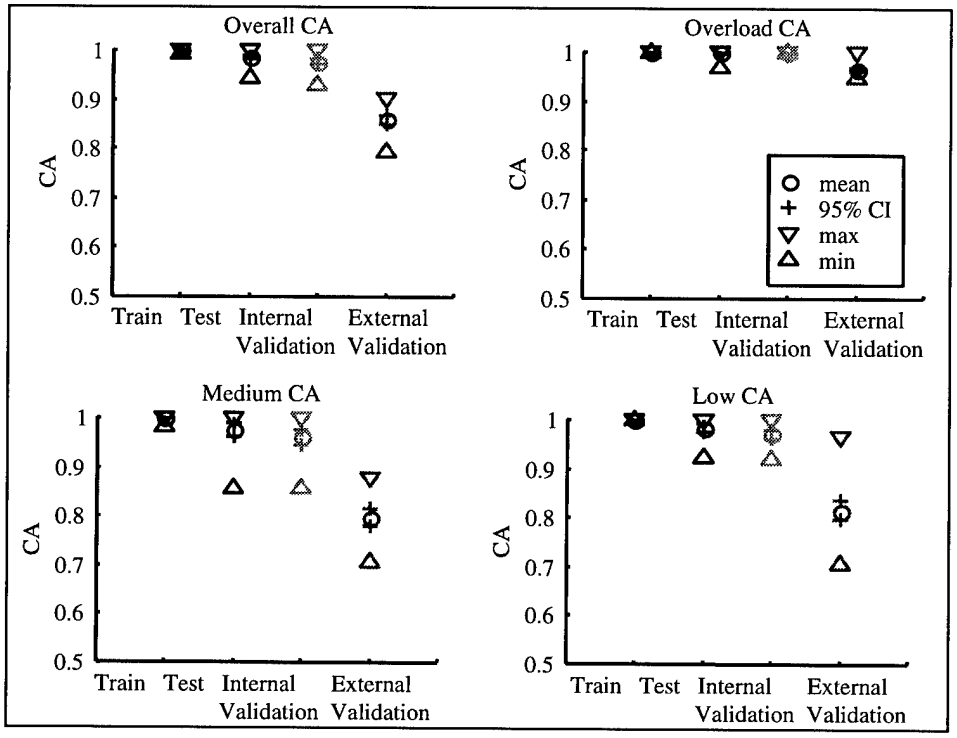

Figure 4-9. Subject 09 Mean CA's.

From the plot above, a couple observations are worth pointing out. First, most CIs for the average CA are within 1% of the calculated mean. Next, the internal validation set appears to be very consistent with the test set CA. Finally, the external validation CA is statistically smaller than the internal validation. In general, while the CA by individual varied greatly, the three patterns pointed out for subject 09 held for most subjects. The overall and individual workload CA confidence intervals were usually within a percent or two of the mean CA, and the test set CA and internal validation set CA were very close with 5 of the 8 95% CI's overlapping. Finally, in all but one case the overall CA decreased from the internal validation set to the external

validation set. A summary of the CA obtained for all eight subjects is contained in Table 4-10.

Table 4-10. CA for All Subjects.

Subject	CI Measure	Train	Test	Internal Validation	External Validation
02	**up 95%**	96.3%	90.2%	88.5%	47.1%
	Mean	**95.8%**	**89.6%**	**87.8%**	**46.5%**
	lo 95%	95.3%	89.0%	87.1%	45.8%
03	**up 95%**	93.6%	84.3%	83.4%	36.2%
	Mean	**93.0%**	**83.5%**	**82.6%**	**35.8%**
	lo 95%	92.5%	82.7%	81.9%	35.3%
04	up 95%	95.3%	84.9%	83.6%	80.2%
	Mean	**94.6%**	**84.0%**	**82.8%**	**79.8%**
	lo 95%	94.0%	83.1%	82.1%	79.4%
05	up 95%	85.3%	80.2%	78.6%	87.7%
	Mean	**84.6%**	**79.5%**	**77.8%**	**87.1%**
	lo 95%	84.0%	78.7%	77.0%	86.5%
09	up 95%	100.0%	98.6%	98.5%	86.1%
	Mean	**100.0%**	**98.4%**	**98.3%**	**85.6%**
	lo 95%	99.9%	98.1%	98.0%	85.1%
11	up 95%	92.2%	87.8%	87.4%	67.9%
	Mean	**91.3%**	**86.9%**	**86.6%**	**67.0%**
	lo 95%	90.5%	86.1%	85.7%	66.1%
13	up 95%	81.4%	74.0%	71.1%	48.4%
	Mean	**80.0%**	**72.8%**	**70.1%**	**47.5%**
	lo 95%	78.5%	71.6%	69.0%	46.6%
16	up 95%	99.0%	95.0%	94.3%	81.4%
	Mean	**98.7%**	**94.6%**	**93.8%**	**81.0%**
	lo 95%	98.4%	94.1%	93.3%	80.6%

4.4 Modeling All 8 Subjects

This final section of the Chapter 4, presents the methodology used to train the linear models, MLP ANNs, and to select an optimal set of features with all individuals used as one collective data set. First, two discriminant models were trained with feature reduction and rank ordering by coefficients and loadings. Next, SNR screening was

performed multiple times using all test subjects as input to one ANN. Finally, the results of the above techniques were analyzed to determine optimal sets of universal input variables.

4.4.1 Linear Group Models. The method used to perform discriminant analysis and linear feature screening was identical to that used for an individual. The only difference was the composition of training and validation data sets. All eight subject's normalized training data was used to create a group training set used to define model parameter's and facilitate feature screening. Each subject's validation data was added together to create a group validation set. Figure 4-10 and Figure 4-11 show classification accuracy as features are removed using loadings and coefficients. From these graphs, classification accuracy between 50% and 60% appears consistent for the three-class problem, which is definitely better than chance (33%). Additionally, unlike the individually trained discriminant models, the CA for the medium workload condition was actually classified correctly a significant portion of the time. Also, as was seen in the individually trained discriminant models, the CA for the overload condition remains relatively high. In fact, as features are reduced, more of the overload observations are correctly classified, albeit at the expense of misclassifying the low and medium conditions. From these two graphs, it is difficult to determine how many features to include as salient if looking at the overall CA. Although if the point of the reduced CA for the medium observations is used as a guide, both models appear to lose predictive power when only 10 features are left in the model.

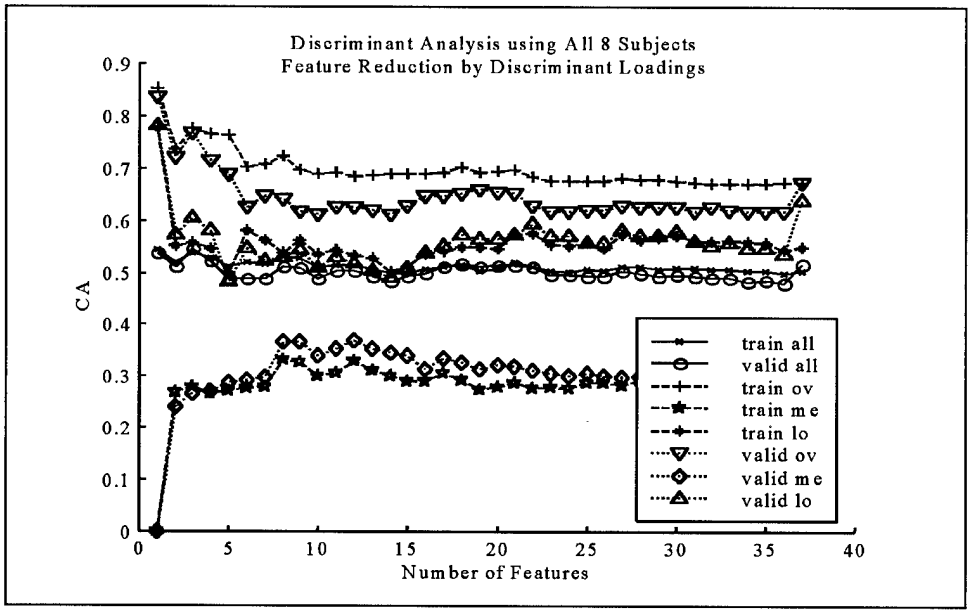

Figure 4-10. Group CA with Feature Reduction by Loadings.

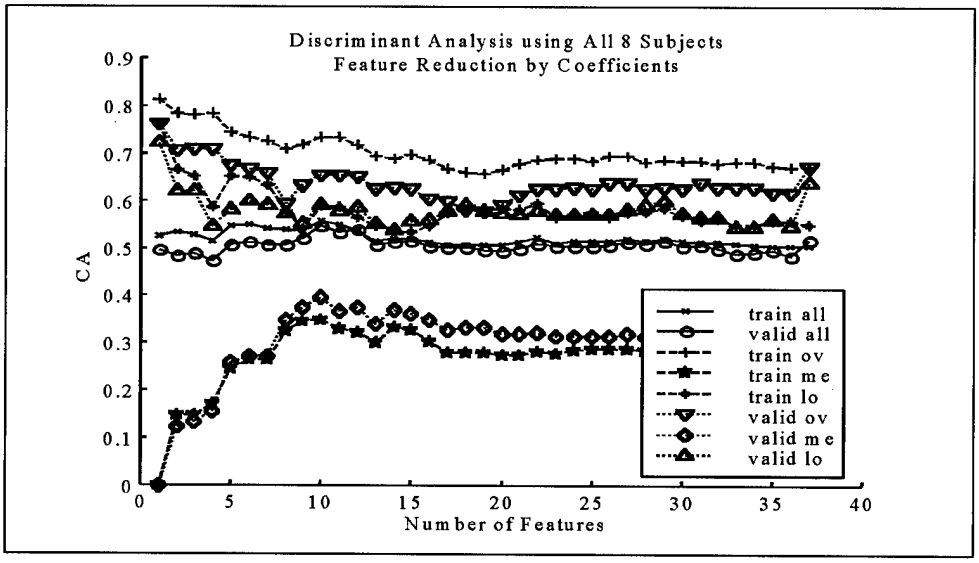

Figure 4-11. Group CA with Feature Reduction by Coefficients.

4.4.2 Group SNR Feature Screening. The method used to perform group SNR screening was identical to that used for individual SNR screening except for the

composition of the training, test, and validation sets. Now, to create the homogenous randomized set of training and test data, each of the eight subject's training data was randomly permuted with the first half assigned to the global training set and the second half assigned to the global test set. The validation set remained a composition of all subject's validation blocks as used for the discriminant modeling. Similar to the individual subject SNR screening, three architectures were used with the SNR screening including 18, 36, and 72 hidden nodes. Of some interest is the computational time involved. While an individual SNR screening run required approximately 1 hour to complete using a PC with a 266 MHz Pentium II processor, the equivalent group SNR screening run required approximately 9.5 hours. Thus, the training time increased almost directly in proportion to the increase in the number of training exemplars. Similar ratios were also experienced with ANNs containing 18 and 36 hidden nodes.

Figure 4-12 and Figure 4-13 are representative of the SNR training. Figure 4-12 shows overall CA for training, test and validation sets. It is further broken down by workload class. Of particular interest is that the ANN medium CA is much better than that seen in either linear model. Additionally, the medium CA appears robust to feature reduction. Low and overload classification definitely shows significant decreases at some point. This point appears to be approximately 10 features for overload, and approximately 20 features for the low workload classification. Also, if test set CA is compared to the linear model training set CA, or the CA for the two validation sets is compared, the ANN can be seen to perform markedly better than the linear models. This increase is approximately 10% on average.

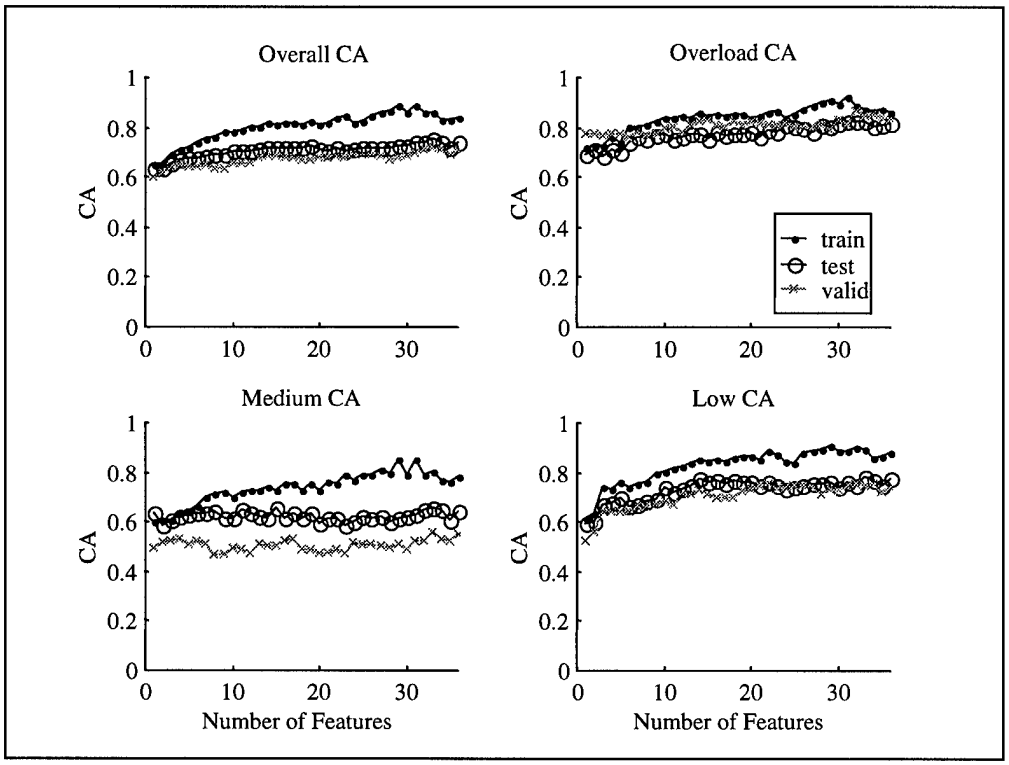

Figure 4-12. Group SNR CA by Workload.

As a final note, a couple interesting trends can be seen in Figure 4-13. First, some subjects consistently perform better than the mean. Subjects 09, 11, and 16 represent this group of subjects. Additionally other subjects can be seen to perform consistently below the mean. Subjects 02, 05 and 13 represent this group, in which subject 13 is clearly the outlying subject that is difficult to classify. After reviewing individual subject CA obtained using individually trained ANNs, a potential pattern was beginning to emerge. Subjects who were modeled well with individually trained models were modeled well in a group model, and subjects who modeled poorly in individually trained models modeled poorly in group trained models.

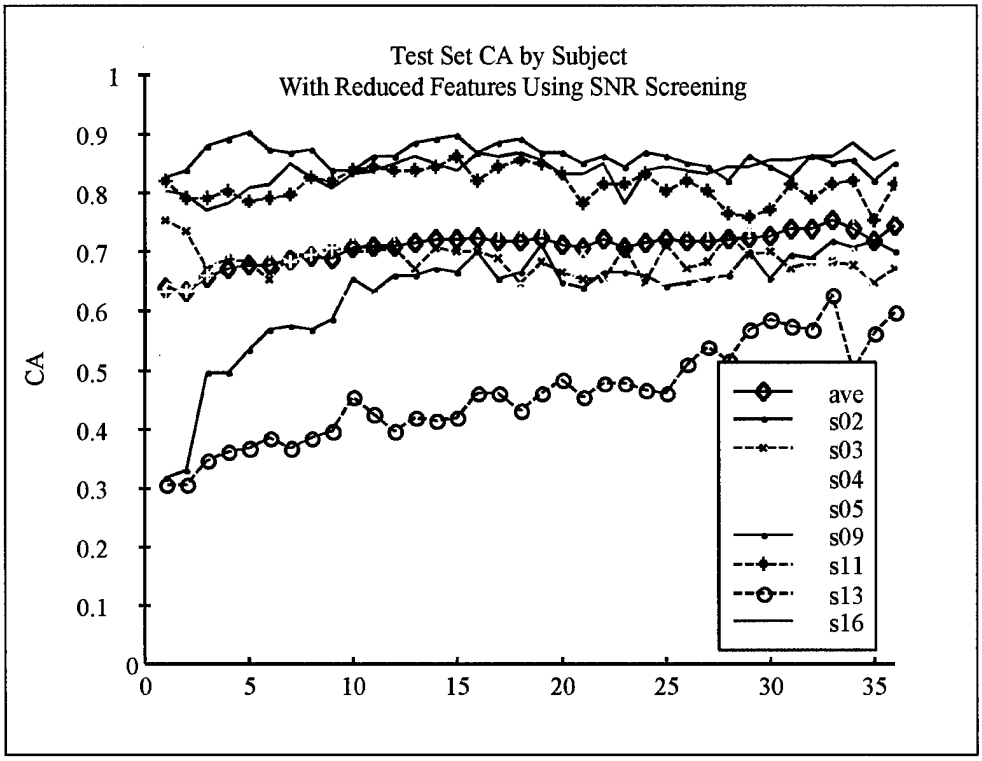

Figure 4-13. Group SNR CA by Subject.

4.4.3 Group Feature Selection. Using the CA plots and corresponding list of features by saliency, the goal now was to determine if a global set of salient features appeared for all eight test subjects. Table 4-11 summarizes the features as ranked by the linear screening and SNR ANN screening. For the group SNR screening, 5 total runs were performed with two runs when using 18 and 36 hidden nodes. As was done for subject 09, all ultrabeta features and the two eye blink features are highlighted in bold. This set of eight features, appears consistently within the top 5 and 10 salient features. In all five SNR screening runs, PZ-ub was selected as the most salient feature. Additionally, EEG beta features appear to be the next most salient set of features, while delta and theta EEG features appear least salient.

Table 4-11. Features by Group Saliency.

Rank	Linear Loading	Linear Coefficien	SNR 18 Nodes	SNR 18 Nodes	SNR 36 Nodes	SNR 36 Nodes	SNR 72 Nodes
1	**O2-ub**	**T7-ub**	**PZ-ub**	**PZ-ub**	**PZ-ub**	**PZ-ub**	**PZ-ub**
2	**T8-ub**	**F7-ub**	**T7-ub**	**T7-ub**	**O2-ub**	HR	**T7-ub**
3	T8-b	**PZ-ub**	**IBLI**	IBRI	**IBLI**	**IBLI**	**T8-ub**
4	**PZ-ub**	**T8-ub**	**F7-ub**	**IBLI**	**T8-ub**	IBRI	**blnks**
5	**IBLI**	**O2-ub**	PZ-b	**blnks**	**F7-ub**	**F7-ub**	PZ-b
6	T7-b	**IBLI**	T7-a	F7-b	HR	**T7-ub**	O2-b
7	**T7-ub**	T7-b	O2-b	O2-a	PZ-b	O2-b	IBRI
8	FZ-b	FZ-b	F7-b	T7-t	FZ-b	**FZ-ub**	**O2-ub**
9	**F7-ub**	HrVar	**O2-ub**	**FZ-ub**	F7-b	T7-a	**F7-ub**
10	**blnks**	PZ-b	T7-t	**O2-ub**	T7-b	**blnks**	T7-a
11	**FZ-ub**	F7-a	HR	FZ-b	**blnks**	F7-b	T8-a
12	O2-b	T7-a	**FZ-ub**	**F7-ub**	brths	PZ-b	T8-b
13	FZ-d	F7-b	FZ-t	F7-a	O2-b	F7-d	F7-b
14	T8-d	FZ-a	F7-d	F7-d	T7-a	FZ-t	FZ-t
15	FZ-t	T7-t	F7-t	FZ-t	F7-d	**T8-ub**	HR
16	FZ-a	FZ-t	T8-t	PZ-b	FZ-t	F7-t	O2-a
17	HrVar	T8-t	**blnks**	O2-t	T8-a	T8-b	T7-b
18	O2-d	O2-a	FZ-d	T8-t	PZ-t	PZ-t	FZ-b
19	PZ-t	PZ-a	O2-t	FZ-d	**T7-ub**	T7-t	PZ-t
20	O2-t	F7-t	O2-a	T7-d	**FZ-ub**	T8-t	O2-d
21	O2-a	**blnks**	PZ-d	HR	PZ-a	F7-a	**IBLI**
22	F7-t	**FZ-ub**	FZ-a	HrVar	F7-t	FZ-b	FZ-a
23	PZ-b	FZ-d	O2-d	T7-b	O2-t	T8-a	F7-a
24	brths	O2-d	PZ-a	T8-d	T7-d	O2-a	F7-t
25	F7-d	PZ-d	brths	T7-a	PZ-d	PZ-d	T8-d
26	F7-a	T8-b	T8-a	O2-b	IBRI	FZ-d	PZ-a
27	T7-d	brths	**T8-ub**	PZ-t	F7-a	**O2-ub**	O2-t
28	T7-t	O2-b	T7-b	FZ-a	O2-a	brths	brths
29	T8-a	T8-d	IBRI	brths	FZ-a	T7-d	T8-t
30	PZ-d	F7-d	T8-b	PZ-d	T8-b	PZ-a	**FZ-ub**
31	T7-a	O2-t	HrVar	PZ-a	T8-t	FZ-a	FZ-d
32	HR	IBRI	T8-d	F7-t	HrVar	O2-d	T7-t
33	F7-b	T7-d	FZ-b	T8-b	T7-t	O2-t	HrVar
34	IBRI	T8-a	PZ-t	**T8-ub**	O2-d	T8-d	F7-d
35	PZ-a	PZ-t	F7-a	T8-a	FZ-d	HrVar	T7-d
36	T8-t	**noise**	T7-d	O2-d	T8-d	T7-b	PZ-d
37	**noise**	HR					

Using the CA plots as a guide, four separate parsimonious sets of salient features were selected. These sets contained 2, 5, 10, and 15 features. Each set was built upon the last. As was true for the selection of individual features, this was performed in a

heuristic manner that does not guarantee the set selected to be the one best set containing a specified number of features. For example, when selecting the top 2 features, PZ-ub was clearly to be included, but the second feature was not as clear. Because, O2-ub may be highly correlated to PZ-ub due to close spatial proximity it was not selected. Using a similar train of thought, no ultrabeta feature was selected, rather IBLI was added as the second feature in an attempt to capture an independent factor to augment PZ-ub's predictive ability. Table 4-12 summarizes the top features selected.

Table 4-12. Top Global Features.

Global Top Features				
Rank	Top 2	Top 5	Top 10	Top 15
1	PZ-ub	PZ-ub	PZ-ub	PZ-ub
2	IBLI	IBLI	IBLI	IBLI
3		02-ub	02-ub	02-ub
4		T7-ub	T7-ub	T7-ub
5		T8-ub	T8-ub	T8-ub
6			F7-ub	F7-ub
7			blnks	blnks
8			O2-ub	O2-ub
9			F7-b	F7-b
10			IBRI	IBRI
11				PZ-b
12				HR
13				FZ-ub
14				O2-a
15				T7-a

In order to assess the feasibility of determining if one net could fit all and to determine is a single set of group features could provide robust results when compared to individually trained ANNs with individually selected input features, ANN models were trained in one of two fashions. First, an individual ANN was trained 30 times for each

individual using each of the four sets of identified group features. All ANN parameter settings were kept as described in Section 4.3. Next, four group nets were trained 30 times with the four different sets of global features. As performed in group SNR training, each individual's 30-minute period of training data was randomly permuted. But now, the random permution was performed 30 times. Thus, for each individual, a total of nine estimates of mean CA were made for training, test, internal-validation, and for the external-validation set. Five of the nine estimates were from individually trained ANNs with the other four resulting from the group ANNs. One group set of exemplars was used to train and test the group nets. After training each individual's training, training-test, and validation sets were presented to the ANN to calculate the appropriate CA by individual. The results of these runs are presented in Chapter 5.

V. Results and "One Net" Methodology

This chapter includes a summary of the results of the classification efforts completed to date. Included is a quick review of the CA achieved using the linear discriminant models. This is followed by a comparison of the CA achieved by subject using three groups of ANNs: individually trained using individual features, individually trained using group features, and group trained using the group features. After the initial results are presented, the salient features and experimental layout are analyzed to determine possible causes of low CA. Finally, the methodology used to determine if, "one net can fit all," is presented along with the associated results.

5.1 Initial Results

The initial results will concentrate on the classification accuracy (CA) achieved by various models. Additionally, these results will be used to assess whether or not a parsimonious set of salient group features appears feasible.

5.1.1 Discriminant Models. As a whole, all discriminant models performed poorly when attempting to classify all three workload groups. This statement is based on the fact that a CA significantly different than zero was only achieved by one of the eight individuals. On the other hand, CA for low and overload remained relatively high. These were both typically 90% to 100% when using some subset of the original features. In contrast to the individual subject models, with enough features, the two group models did classify some of the medium workload correctly. Yet, the medium classification was typically no better than chance (33%), and diminished quickly when less than 10 features

where left in the model. Additionally, the group discriminant models maintained fairly stable test and validation set CA, regardless of the number of features used.

Estimates of the optimal CA achieved by individuals and the group using loading and coefficient feature reduction were obtained by selecting the best number of features from the plots of CA. As with all feature selection in this effort, an optimal set appeared to have the minimal number of features with a relatively high and robust CA. Table 5-1 summarizes the CA obtained by subject and feature removal method. Note, a CA of 66% or 67% is indicative of all low and overload observations being correctly classified with all medium observations misclassified.

Table 5-1. Discriminant Model CA.

Discriminant Analysis Classification Accuracy				
	Feature Reduction by Loading		Feature Reduction by Coefficient	
Subject	Training CA	Validation CA	Training CA	Validation CA
02	60%	60%	53%	53%
03	67%	67%	67%	67%
04	62%	62%	62%	62%
05	66%	68%	66%	70%
09	65%	65%	67%	67%
11	67%	67%	67%	67%
13	65%	65%	65%	65%
16	67%	67%	66%	66%
Group	55%	55%	55%	55%

5.1.2 Individual and Group ANNs Two primary methods were used to compare classification models. For each of these methods, CA was used as the measure of effectiveness. For each individual, comparisons between CA for any combination of nine ANN models can be performed. The first model presented is the individually trained ANN using individually selected features. This model was used as a basis for the comparison of the remaining eight models. The next four models were individually

trained using the top 2, 5, 10, and 15 group features. The last four models were derived using the four group trained ANNs, when only presenting one subject's data to the trained ANN.

In addition to comparing the CA of the previously assigned "external" validation sets, an "internal" validation set was formed using 25% of the training exemplars. As before, 50% of the training exemplars was assigned to the training set. But now, only 25% of the training exemplars was assigned to the training-test set. The last 25% of the training exemplars was then assigned as the internal validation set. Results for the training-test set and internal validation set were very consistent. For all models, the overall CA means for test set and internal validation were statistically equivalent, including similar values of standard error. This suggests that enough exemplars were included in the training and test sets to use the test set CA as an unbiased estimator for the model's CA for observations homogeneous to the training set.

The first method of evaluation uses the mean values for each CA and computation of 95% confidence intervals (CI) about the means. To facilitate comparisons, the CI for the CA obtained in the first model, is used as a statistic to test the remaining eight CA CIs against. In Table 5-2, if the CI for CA overlaps or is greater than the CI for the individually trained model using individual features, the associated mean is highlighted in bold print. This suggests that the model is statistically equivalent or better than the individually trained ANN using individually selected features. On the other hand, if the CI about the mean is significantly less than the CA obtained when using the individually trained ANN with individually selected features, the mean value is not highlighted. This

represents a model that does not appear to perform as well as the first ANN for each subject.

From Table 5-2, a few general observations can be made. As mentioned, to facilitate comparisons, the CA CI for the individually trained ANNs using individually selected features is presented first for each subject. The CA CIs for the four individually trained ANNs using group features are presented next on the same line. Finally, below the four individually trained ANNs, the group trained ANN CA CIs are presented. As may be expected, the CA obtained for the internal validation is much better than that obtained for the external validation sets. This is expected, because the ANNs are trained, tested, and internally validated on a random sample of data from the same 30-minute period. Additionally, for the internal validation sets, the individually trained ANNs perform better than the group trained ANNs. Again this is expected as the group ANNs train using training and test sets composed of all eight individuals, in which person specific patterns are less likely to emerge.

When comparing the external validation sets, a different pattern of results can be seen. For six of the eight subjects, the CA obtained using the group ANNs and features appeared equivalent to, or better than, the individually trained ANN using individually selected features. In fact, when validating on an external block of data, the same six subjects obtained improvement in overall CA when using one of the four group ANNs for workload classification. While this is encouraging for the use of one net and one set of salient features, the overall CA remained fairly low for some of these subjects. Thus, to gain insight into the overall CA reported by each model, confusion matrices were analyzed next.

Table 5-2. CA CIs by Subject.

Sub 02	Individual features			Top 2 Group Features			Top 5 Group Features			Top 10 Group Features			Top 15 Group Features		
	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%
In-Valid	87.1%	**87.8%**	88.5%	61.5%	**63.6%**	65.7%	72.9%	**74.4%**	75.8%	79.5%	**81.1%**	82.7%	87.7%	**89.1%**	90.4%
				43.1%	**45.5%**	47.9%	50.8%	**52.6%**	54.5%	59.7%	**61.8%**	64.0%	71.8%	**73.4%**	75.1%
Ex-Valid	45.8%	**46.5%**	47.1%	61.0%	**62.9%**	64.8%	58.4%	**60.6%**	62.7%	70.4%	**72.0%**	73.7%	48.2%	**50.2%**	52.3%
				43.1%	**45.5%**	47.9%	50.8%	**52.6%**	54.5%	59.7%	**61.8%**	64.0%	71.8%	**73.4%**	75.1%
Sub 03	Individual features			Top 2 Group Features			Top 5 Group Features			Top 10 Group Features			Top 15 Group Features		
	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%
In-Valid	81.9%	**82.6%**	83.4%	77.6%	**79.2%**	80.8%	79.5%	**80.8%**	82.1%	80.2%	**82.1%**	83.9%	80.3%	**82.1%**	83.9%
				70.4%	**73.0%**	75.5%	69.5%	**71.1%**	72.8%	72.0%	**73.5%**	75.0%	69.9%	**71.6%**	73.2%
Ex-Valid	35.3%	**35.8%**	36.2%	48.5%	**49.8%**	51.2%	44.1%	**45.3%**	46.5%	34.4%	**35.5%**	36.6%	34.2%	**35.0%**	35.8%
				70.4%	**73.0%**	75.5%	69.5%	**71.1%**	72.8%	72.0%	**73.5%**	75.0%	69.9%	**71.6%**	73.2%
Sub 04	Individual features			Top 2 Group Features			Top 5 Group Features			Top 10 Group Features			Top 15 Group Features		
	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%
In-Valid	82.1%	**82.8%**	83.6%	59.6%	**62.8%**	65.9%	71.4%	**73.0%**	74.6%	77.8%	**79.1%**	80.5%	81.9%	**83.1%**	84.3%
				62.5%	**64.3%**	66.2%	66.1%	**67.7%**	69.4%	70.8%	**72.3%**	73.9%	74.0%	**75.6%**	77.2%
Ex-Valid	79.4%	**79.8%**	80.2%	55.4%	**58.4%**	61.3%	70.1%	**71.9%**	73.6%	74.5%	**75.5%**	76.6%	75.3%	**76.2%**	77.2%
				62.5%	**64.3%**	66.2%	66.1%	**67.7%**	69.4%	70.8%	**72.3%**	73.9%	74.0%	**75.6%**	77.2%
Sub 05	Individual features			Top 2 Group Features			Top 5 Group Features			Top 10 Group Features			Top 15 Group Features		
	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%
In-Valid	77.0%	**77.8%**	78.6%	61.9%	**64.4%**	67.0%	81.7%	**83.0%**	84.4%	81.2%	**83.0%**	84.8%	80.1%	**81.4%**	82.7%
				61.9%	**63.8%**	65.7%	64.8%	**66.6%**	68.3%	66.6%	**68.2%**	69.7%	65.6%	**67.1%**	68.6%
Ex-Valid	86.5%	**87.1%**	87.7%	55.9%	**58.2%**	60.5%	82.1%	**83.5%**	84.9%	71.2%	**72.7%**	74.2%	66.1%	**68.1%**	70.2%
				61.9%	**63.8%**	65.7%	64.8%	**66.6%**	68.3%	66.6%	**68.2%**	69.7%	65.6%	**67.1%**	68.6%
Sub 09	Individual features			Top 2 Group Features			Top 5 Group Features			Top 10 Group Features			Top 15 Group Features		
	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%
In-Valid	98.0%	**98.3%**	98.5%	83.2%	**85.7%**	88.3%	93.7%	**94.7%**	95.7%	96.2%	**96.9%**	97.5%	97.0%	**97.5%**	97.9%
				82.7%	**85.1%**	87.5%	85.3%	**86.4%**	87.6%	86.8%	**88.2%**	89.5%	86.8%	**88.1%**	89.4%
Ex-Valid	85.1%	**85.6%**	86.1%	83.4%	**85.1%**	86.8%	75.6%	**76.7%**	77.9%	85.1%	**86.2%**	87.2%	84.7%	**86.1%**	87.6%
				82.7%	**85.1%**	87.5%	85.3%	**86.4%**	87.6%	86.8%	**88.2%**	89.5%	86.8%	**88.1%**	89.4%
Sub 11	Individual features			Top 2 Group Features			Top 5 Group Features			Top 10 Group Features			Top 15 Group Features		
	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%
In-Valid	85.7%	**86.6%**	87.4%	72.4%	**75.5%**	78.6%	83.1%	**84.2%**	85.3%	84.2%	**85.4%**	86.5%	81.8%	**83.3%**	84.8%
				72.3%	**74.4%**	76.5%	76.6%	**78.0%**	79.5%	76.7%	**78.5%**	80.4%	75.8%	**77.4%**	79.0%
Ex-Valid	66.1%	**67.0%**	67.9%	74.1%	**77.2%**	80.3%	72.4%	**73.7%**	75.1%	78.1%	**79.1%**	80.1%	80.0%	**81.0%**	82.1%
				72.3%	**74.4%**	76.5%	76.6%	**78.0%**	79.5%	76.7%	**78.5%**	80.4%	75.8%	**77.4%**	79.0%
Sub 13	Individual features			Top 2 Group Features			Top 5 Group Features			Top 10 Group Features			Top 15 Group Features		
	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%
In-Valid	69.0%	**70.1%**	71.1%	48.7%	**51.0%**	53.2%	61.2%	**62.8%**	64.4%	66.9%	**69.1%**	71.4%	71.9%	**73.6%**	75.4%
				33.4%	**35.3%**	37.1%	40.1%	**42.1%**	44.2%	45.2%	**47.0%**	48.8%	48.5%	**50.6%**	52.7%
Ex-Valid	46.6%	**47.5%**	48.4%	43.9%	**46.5%**	49.1%	31.1%	**32.4%**	33.8%	39.3%	**41.1%**	42.9%	44.2%	**45.8%**	47.3%
				33.4%	**35.3%**	37.1%	40.1%	**42.1%**	44.2%	45.2%	**47.0%**	48.8%	48.5%	**50.6%**	52.7%
Sub 16	Individual features			Top 2 Group Features			Top 5 Group Features			Top 10 Group Features			Top 15 Group Features		
	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%	lo 95%	Mean	up 95%
In-Valid	93.3%	**93.8%**	94.3%	80.3%	**81.6%**	83.0%	87.9%	**88.9%**	89.9%	88.0%	**89.5%**	90.9%	88.1%	**89.5%**	90.9%
				75.7%	**78.4%**	81.1%	77.7%	**79.7%**	81.7%	80.1%	**81.7%**	83.2%	83.3%	**84.9%**	86.5%
Ex-Valid	80.6%	**81.0%**	81.4%	68.1%	**68.7%**	69.4%	80.3%	**81.5%**	82.8%	82.7%	**83.6%**	84.5%	84.1%	**85.0%**	85.9%
				75.7%	**78.4%**	81.1%	77.7%	**79.7%**	81.7%	80.1%	**81.7%**	83.2%	83.3%	**84.9%**	86.5%

Equally important to the overall CA is the composition of correctly and incorrectly classified observations. For example, a model may appear to perform only slightly better than chance, but if the misclassifications are inspected, the model may be

classifying better than chance, but incorrectly. An example of this was seen after analyzing the confusion matrices (CMs) containing all classification results from the 30 ANN runs used to calculate the mean CA values. A sample CM is shown in Table 5-3.

Table 5-3. Sample Confusion Matrix.

		Model Classification			
		Low	**Medium**	**Overload**	**Total**
True Classes	**Low**	Low CA	Error	Error	Total Low Exemplars
	Medium	Error	Medium CA	Error	Total Medium Exemplars
	Overload	Error	Error	Overload CA	Total Overload Exemplars
	Total	Total Classified Low	Total Classified Medium	Total Classified Overload	Overall CA

Within each cell of the CM both the number of correct classifications and the associated percentage of correct classifications is included. Appendix A contains CMs by subject for each of the nine ANN models calculated. The internal validation set CM can be used to assess a model's internal predictive power for observations homogeneous to those in the training set, and the external validation set CM can be used to assess a model's predictive power of exemplars outside of the training set.

While the internal validation set contains a random composition of two blocks of data, the external validation set includs only those observations from one block. Because external validation set observations were from only one block, patterns of

misclassification possibly associated with a specific block effect can be identified. One example of this is presented in Table 5-4 for subject 03.

Table 5-4. Subject 03 Confusion Matrices.

Individually Trained ANN **Internal Validation** 10-features	681	157	0	838	Group Trained ANN **Internal Vaidation** 10-features	610	302	7	919
	81.3%	18.7%	0.0%			66.4%	32.9%	0.8%	
	179	646	70	895		200	576	69	845
	20.0%	72.2%	7.8%			23.7%	68.2%	8.2%	
	2	60	815	877		20	94	732	846
	0.2%	6.8%	92.9%			2.4%	11.1%	86.5%	
	862	863	885	2610		830	972	808	2610
				82.07%					73.49%
Individually Trained ANN **External Validation** 10-features	147	1527	66	1740	Group Trained ANN **External Validation** 10-features	434	1250	56	1740
	8.4%	87.8%	3.8%			24.9%	71.8%	3.2%	
	1468	244	28	1740		1065	601	74	1740
	84.4%	14.0%	1.6%			61.2%	34.5%	4.3%	
	17	261	1462	1740		5	269	1466	1740
	1.0%	15.0%	84.0%			0.3%	15.5%	84.3%	
	1632	2032	1556	5220		1504	2120	1596	5220
				35.50%					47.91%

The CMs are from two ANN models. On the left, the top-10 group features were used as input for an ANN that was individually trained for subject 03, on the left are the results for subject 03 from the group ANN using the same 10-features. Of particular interest here is that while a significant difference in overall CA is observed from internal to external validation CA, the overload CA for both internal and external validation remains well above 80%, even though the overall CA is only 36% for the independently trained ANN. The cause of this phenomenon is quickly observed in the CM. The model is classifying low as medium and medium as low in both of the external validation CMs. Other examples of confusion between low and medium can be seen for subjects 05, 11, 13, and 16, where in at least one of the external validation CMs, CA is 50% or less for the correct class and 50% or better for the incorrect class.

5.1.3 Group Features. After comparison of the external validation set CA presented in Table 5-2 and review of the corresponding confusion matrices in Appendix A, it appears as if the use of a group set of salient features is feasible. A group trained ANN also appears feasible. To statistically test whether individually trained ANNs using individually selected features are better than group trained ANNs using group features, a test hypothesis is posed. The null hypothesis (H_o) is that the individually trained ANNs are equivalent to the group trained ANNs, or $\hat{p}_{\text{Ind trained}} = \hat{p}_{\text{Grp trained}}$, where $\hat{p}$ is an estimated CA probability and the estimated variance (S^2) is then $\hat{p} \bullet (1-\hat{p})$. The alternative hypothesis (H_a) is that the individually trained ANNs perform better than the group trained ANNs, or $\hat{p}_{\text{Ind trained}} > \hat{p}_{\text{Grp trained}}$. A Z-test can then be used to evaluate the null and alternative hypotheses [47]. For this test, all correct classifications for the 30 trained group ANNs using 5, 10, and 15 features were addeded together to provide an estimate of $\hat{p}_{\text{Grp trained}}$. The classification results from the 30 trained ANNs, for the three models defined by input features, were added together as they appeared to be a homogenous representation of how an optimally trained group ANN could perform. All correct classifications for 100 trained individual ANNs were used to provide an estimate of $\hat{p}_{\text{Ind trained}}$. A summary of the hypothesis tests for differences in the total CA and in the overload CA is included as Table 5-5, where Z_o is calculated as,

$$Z_o = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \tag{5-1}$$

From Table 5-5, with $\alpha = 0.05$ and a corresponding 95% level of confidence, for a one-sided test, statistical evidence is not present to support the alternative hypothesis. Individually trained ANNs do not have statistically greater CA. Both total CA and the overload CA for the individually trained ANNs are not statistically greater than the group trained ANNs. Additionally, the CA obtained from the individual or the group models can be seen to be equivalent in practical terms, with differences of less than 1%. Therefore, the use of group features and group trained ANNs appears feasible.

Table 5-5. Individual vs. Group Hypothesis Testing.

H_o: $\hat{p}_{Ind\ trained} = \hat{p}_{Grp\ trained}$ H_a: $\hat{p}_{Ind\ trained} > \hat{p}_{Grp\ trained}$	**Total CA**	**Overload CA**
$\hat{p}_{Ind\ trained}$	66.27%	83.68%
$\hat{p}_{Grp\ trained}$	67.28%	83.47%
$n_{Ind\ trained}$	139,200	46,400
$n_{Grp\ trained}$	125,280	41,760
Z_o	-5.493	0.826
$Z_o > Z_{.05} = 1.645$	No	No
Conclusion	Fail to Reject H_o	Fail to Reject H_o

Figure 5-1 and figure 5-2 summarize some of the supporting evidence. They include representations of subjects with their individually selected salient features highlighted on each head. These features were identified in Chapter 4 and are summarized in Table 4-8. In most cases, a single dark EEG location indicates an μbeta feature was selected as salient by the individual. Additional dark circles indicate another frequency band of EEG (usually alpha or beta) was also chosen as a salient feature from that location. A very light EEG location indicates that theta was chosen as salient, while no differentiation in color indicates that no feature was chosen from that particular

location. A dark left eye indicates that IBLI is salient and a dark right eye indicates that blinks are salient. Finally, a dark nose indicates IBRI was selected and, the presentation of a heart indicates that HR was a salient feature.

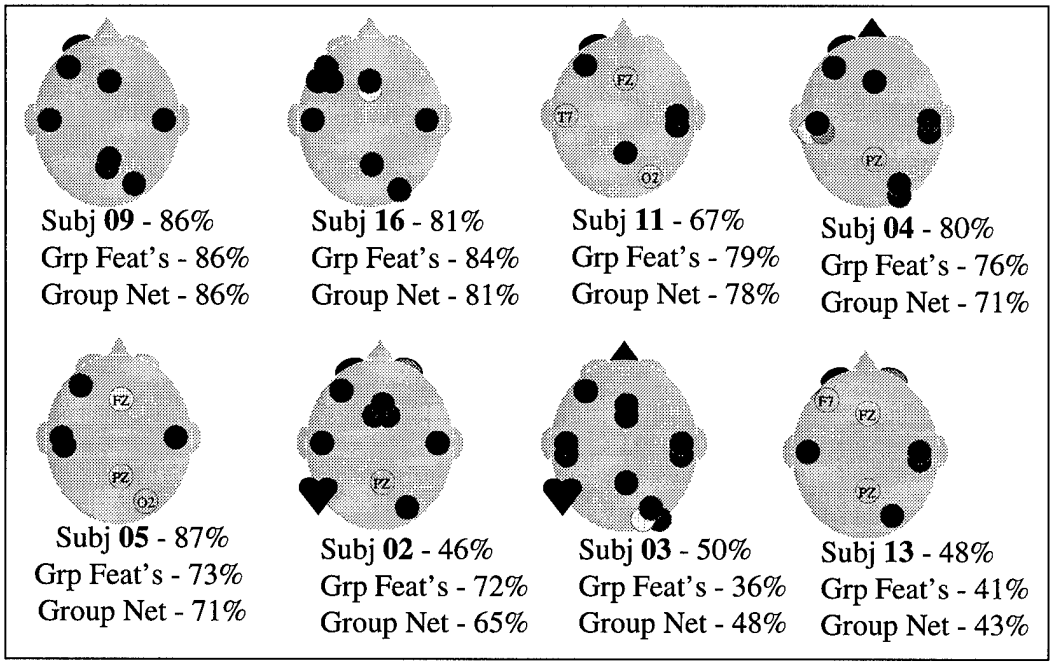

Figure 5-1. Overall CA by Subject.

Figure 5-1 ranks subjects by the CA obtained from individually trained ANNs using individually selected features. Also presented is the CA obtained from individually trained ANNs using the top 10 group features (Grp Feat's) and the group ANN using the top 10 features (Group Net). The top 10 features were used for comparison following inspection of the confusion matrices. After accounting for reversed classification of low and medium, the use of 10 features appeared to be the best of the four group sets of input features. While the overall external validation set CA varies from 36% to 86% depending on subject and ANN model, analysis of the overload classification faired much better.

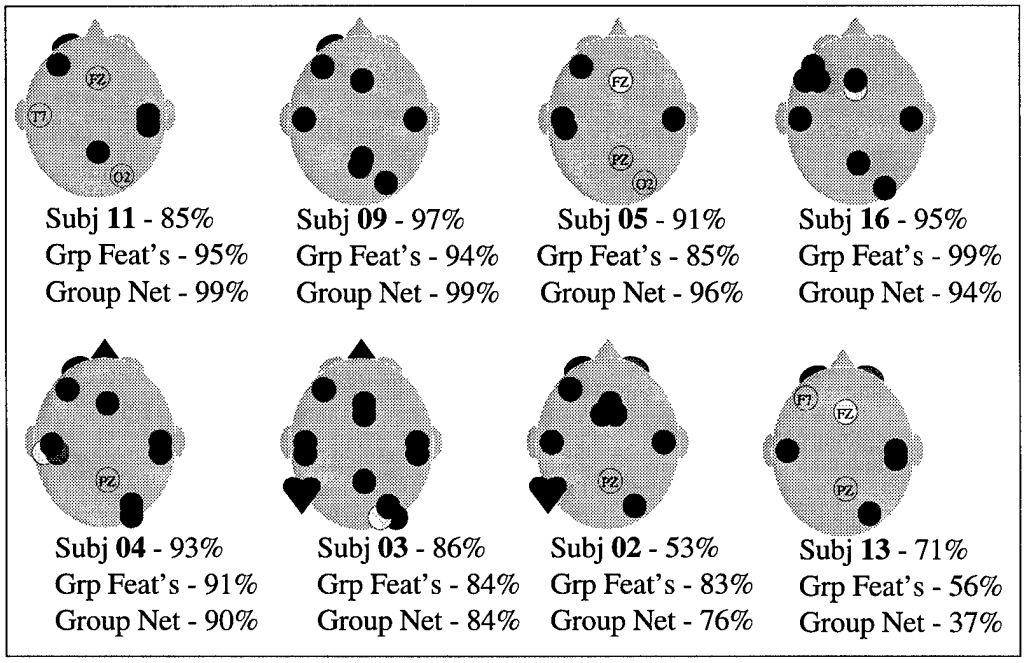

Figure 5-2. Overload CA by Subject.

Figure 5-2 contains the subjects ordered by the CA obtained for overload using the group ANN with 10 input features. Of particular interest is that, in general, each CA presented by subject and model shows improvement when going from the overall CA to just the overload CA. Individually trained ANNs using the group features also appeared to perform just as well as the individually trained ANNs using individually selected features as presented on the first and second lines below each subject. Finally, for all but one subject, the group ANN provided a CA by subject as good as, or better, than the first CA.

In summary, all three models all performed similarly. The average CA across subjects using individually trained ANNs with individually selected features was 84%. The average CA across subjects using individually trained ANNs with the top 10 group features was 86%. And, the average CA across subjects using a group trained ANNs

with the top 10 group features was 84%. The confusion matrices also provided insights into where the models may be having difficulties. To gain insight as to why the low and medium classes were being reversed in some cases, the salient features were analyzed.

5.2 Salient Feature Analysis

As was discovered in the confusion matrices, after training models using 30-minutes of data (2 blocks), the classification of the external 15-minute block of data sometimes resulted in the reverse classification of low and medium workloads. The first step taken to gain insight of this phenomenon was to examine plots of the data including the means by workload level.

5.2.1 Salient Feature Mean Values. To gain possible insight as to why the models performed poorly, the raw data was analyzed using plots that included all of the normalized input values and the associated means by workload level. For these plots, two subjects were plotted at a time. For many features, subject 09 was used as a standard to compare against. Subject 09 was selected as the standard because previous data snooping showed consistent ordering of feature means by workload (either increasing or decreasing with workload depending on feature). Additionally, to date, subject 09 consistently retained high CA in all modeling efforts. As a starting point, PZ-μbeta was plotted for all subjects. PZ-μbeta was selected as it was identified by all five SNR feature saliency runs as the most salient feature for use by ANN. To facilitate the detection of patterns and to compare two subjects on the same plot, all workload data was artificially ordered from low to overload in each of the three blocks. Thus, the actual order of the workload levels presented to each subject did not occur in the following order. Finally,

subject 03 is the primary interest in Figure 5-3 as indicated by the darker plots of mean values and normalized data. Subject 09 is presented in a lighter grayscale as a reference.

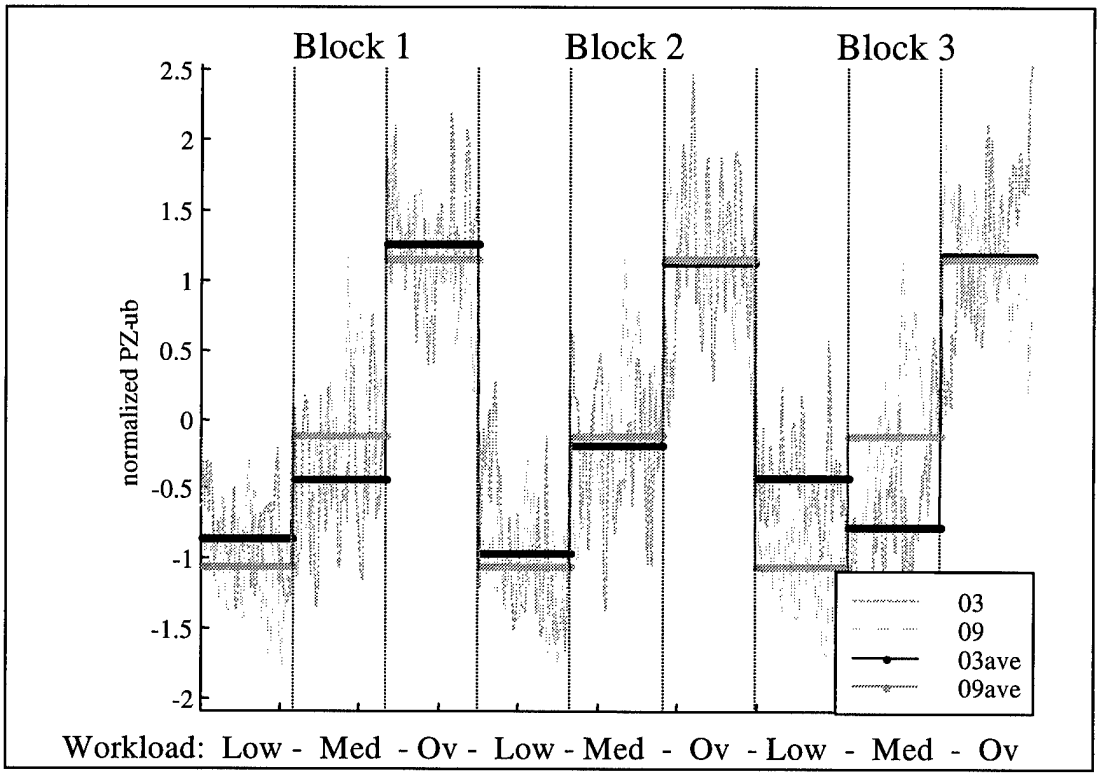

Figure 5-3. Subject 03 vs. 09 PZ-ub.

After reviewing Figure 5-3, and additional plots of the salient features, the cause of low and medium reversal can be traced directly to the input features. For subject 03, the first two blocks that provided the training set data contain a clear pattern of increasing PZ-μbeta as workload increases. The third block, that was used as the external validation set for subject 03, does not have the same pattern. To gain insight as to why this may have occurred, the actual ordering of the workloads within block were analyzed next as will be presented in the following section.

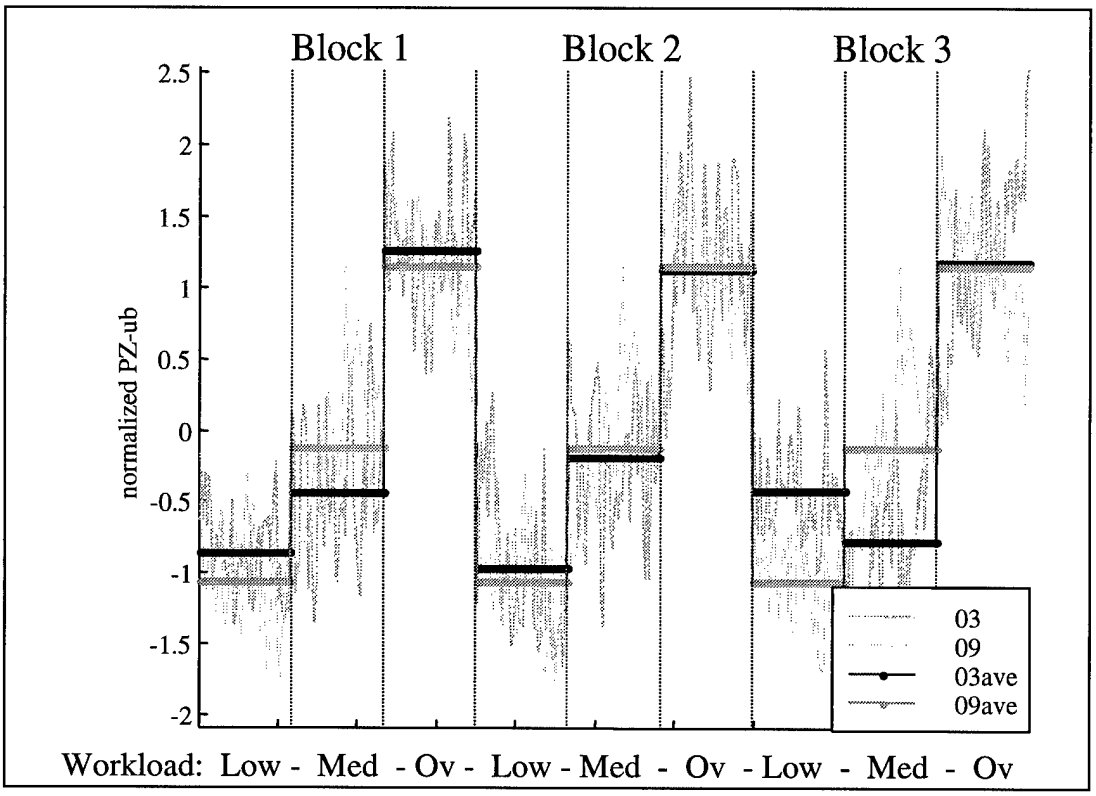

Figure 5-4. Subject 13 vs. 09 PZ-ub.

Figure 5-4 includes normalized PZ-μbeta for subjects 13 and 09. The data for subject 13 is presented as evidence to why CA remained relatively low in the many modeling efforts attempted. In Figure 5-4 the same distinct step pattern is displayed in light gray for subject 09. In contrast, it is difficult to detect any pattern for subject 13. Additionally, while mean values by workload are spaced at about –1, 0, and 1 for subject 09, the means are all within –0.5 and 0.5 standard deviations for subject 13.

As a final note, the variability within each workload level appears much greater for subject 13. To complicate the classification of workload, EEG μbeta data for subject 13 was found to have inconsistent ordering of mean values, small separation of means, and large variance. If an ANN model were primarily using linear logic, it would be very

difficult to correctly classify observations. To validate that the data was correctly identified as low, medium, and overload, subject 13's most salient feature was analyzed. Among the noise and inconsistent means of the EEG data, eye-blink data did appear consistent for subject 13, at least in the ordering of mean values, suggesting that the data was labeled correctly.

Overall, after analysis of PZ-μbeta and other features for each individual, four of the eight subjects were found to have inconsistent orderings of mean input values by workload. In addition, for three of these subjects, at least one block of data contained a 5-minute period of overload work with a lower mean than one of the medium or low workload levels. Thus, the next step of investigation is the analysis of the experimental sequence of workload presentation.

5.2.2 Salient Feature Temporal Effects. To gain insight into the poor classifications and reverse classification of low and medium the graphs of normalized input data were correlated to the actual workload presentation orders presented in Table 5-6. In every case of inconsistent workload means, a temporal effect from the experimental sequence could be hypothesized to explain the inconsistency. For example, in one case, the second of two consecutive medium workload means appeared as big as an overload workload mean. This is an example of the EEG μbeta signal being nonstationary, where it appears to drift-up or increase during equivalent workloads. Other observed temporal effects include the μbeta EEG signals drifting-down or getting smaller during low workload and moving up or down depending on individual and previous workload for medium workload levels. Also, as was specifically seen for subject 03 in block 3, the low workload mean appeared as a medium level. As can be seen in Table 5-6, the low

workload followed a period of overload. One possible explanation is the time required for the EEG μbeta signal to drift back down to a low level was not sufficient for this individual in this particular circumstance. Table 5-6 is presented next.

Table 5-6. Workload Levels by Subject.

Subject	Block 1	Block 2	Block 3
02	Me-Lo-Ov	Ov-Lo-Me	Lo-Ov-Me
03	Ov-Lo-Me	Lo-Ov-Me	Me-Ov-Lo
04	Lo-Ov-Me	Me-Ov-Lo	Ov-Me-Lo
05	Me-Ov-Lo	Ov-Me-Lo	Lo-Me-Ov
09	Ov-Lo-Me	Lo-Ov-Me	Me-Ov-Lo
11	Me-Ov-Lo	Ov-Me-Lo	Lo-Me-Ov
13	Lo-Me-Ov	Me-Lo-Ov	Ov-Lo-Me
16	Lo-Ov-Me	Me-Ov-Lo	Ov-Me-Lo

From Table 5-6, a circle indicates the low period preceded by an overload period for subject 03. Additionally, the last low workload for subject 13 was preceded by two overload periods. Because the μbeta EEG data appeared inconsistent for subject 13 and because subject 13 was presented with back-to-back overload conditions prior to the low, further evidence is found for a correlation between workload presentation order and μbeta EEG inconsistencies.

5.3 One Net Methodology

This chapter has presented evidence supportive of a single group set of salient features. Evidence was also found that low and medium workload was not linearly separable, in which supportive evidence was presented as the reversal in mean values of

low and medium by subject 03. The presentation order of workload appears to be correlated in some fashion to the inconsistencies observed in the normalized data. Therefore, an experiment specifically designed to minimize the temporal effects of the experimental sequence will be presented in the attempt to identify if "one net can fit all." Because a potential application of this research includes the detection of workload as a pilot to transitions from a "nominal" to an "overload" state, the one net methodology will use a corresponding selection of the data. To perform "nominal" vs. "overload" classification, the low and medium classes can be combined to form the nominal class. The grouping of classes in this manner will remove any potential misclassifications between low and medium. The methodology specifics are described in the following section.

5.2.1 Data Selection. "One net" efforts were developed with concentration on the detection of operator performance degradation. In doing so, a two-group classification problem will be attempted. As the ultimate validation test, models will be validated using the data from subjects not used for training.

The first step in the "one net" methodology involves selection of a subset of the original data. For each individual a nominal period followed by an overload period is desired. To minimize potential temporal effects of workload presentation orders, an extended period of nominal workload prior to the nominal and overload period to be used was desired. Additionally, to keep the design balanced, half of the nominal periods were selected as low with the other half selected as medium. Table 5-7 highlights the data to be used.

Table 5-7. "One Net" Data Set Selection.

Subject	Block 1	Block 2	Block 3
02	Me-Lo-Ov	Ov-Lo-Me	Lo-Ov-Me
03	Ov-Lo-Me	Lo-Ov-Me	Me-Ov-Lo
04	Lo-Ov-Me	Me-Ov-Lo	Ov-Me-Lo
05	Me-Ov-Lo	Ov-Me-Lo	Lo-Me-Ov
09	Ov-Lo-Me	Lo-Ov-Me	Me-Ov-Lo
11	Me-Ov-Lo	Ov-Me-Lo	Lo-Me-Ov
13	Lo-Me-Ov	Me-Lo-Ov	Ov-Lo-Me
16	Lo-Ov-Me	Me-Ov-Lo	Ov-Me-Lo

Once the data was selected, a methodology was then developed for determining the number of models required and the subjects to be used as validation for each model. Using the group ANN overload CA as a measure of potential CA for each subject, all eight subjects were then ordered. The first decision was to remove the "best" and "worst" subjects for use in training a group model. As can be seen in Figure 5-5, subjects 09 and 13 were separated from the group and will be used for validation purposes only.

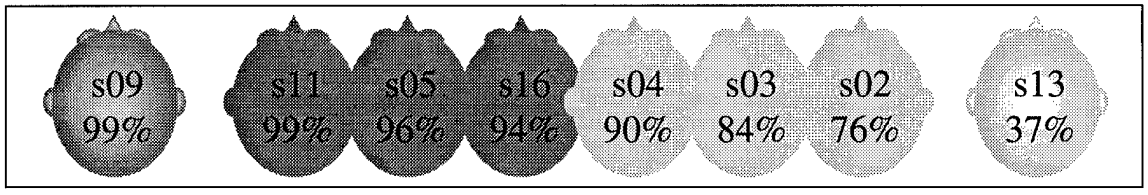

Figure 5-5. Subjects by Group Net Overload CA.

With six subjects remaining, a decision was made to train three separate models for the group, using two of the six subjects for model validation in addition to subjects 09 and 13. The use of three models in this manner will provide a validation set CA for each of the subjects. For six subjects only one model will determine CA, while for the "best"

and "worst" subjects, CA can be computed as their average of all three models. To maintain a balanced experiment, one subject with low to overload data and one subject with medium to overload data were selected as the additional validation subjects for each of the three model. The remaining four subjects were then used to form a training and a training-test set. Randomly permuting the three low to overload and the three medium to overload subjects, validation and training set assignments were made as indicated in Table 5-8.

Table 5-8. "One Net" Training and Test Sets.

	Training and Test		Validation Sets			
Model	Lo to Ov	Me to Ov	Lo to Ov	Me to Ov	Lo to Ov	Me to Ov
1	s02, s04	s11, s16	s03	s05	s13	s09
2	s02, s03	s05, s11	s04	s16	s13	s09
3	s03, s04	s05, s16	s02	s11	s13	s09

5.2.2 Linear Modeling. The first step in this reduced two-group classification effort involved using linear discriminant models. These models were performed to assess linear CA, provide a benchmark for a ANN CA, and to provide insight into salient feature selection. The linear methodology included performing six saliency screening runs as were described in Chapter 4 for three classes. Thus, feature reduction was performed using both loadings and coefficients for each of the three "one net" models as defined by the structure subjects used for validation and training sets. Samples of two of the six models are provided in Figure 5-6 and Figure 5-7.

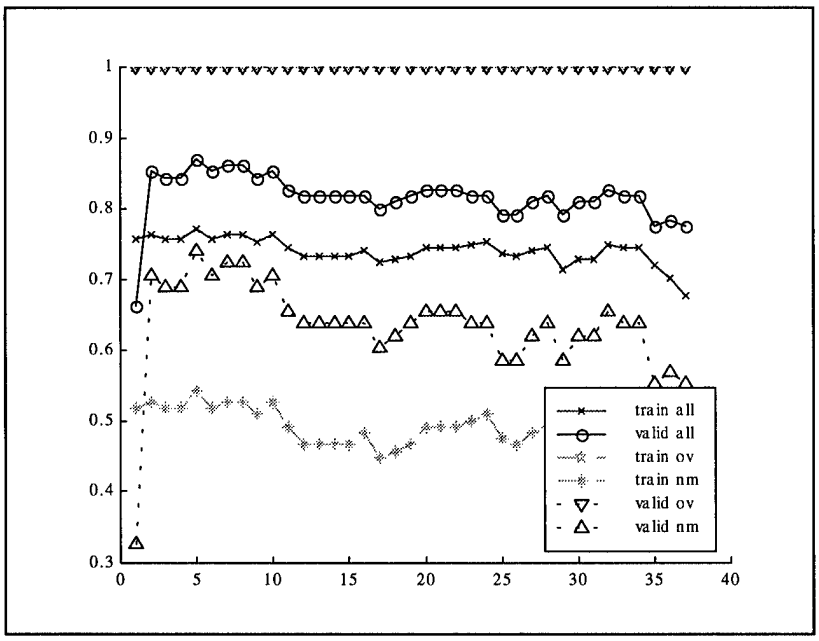

Figure 5-6. CA with Feature Reduction by Loading.

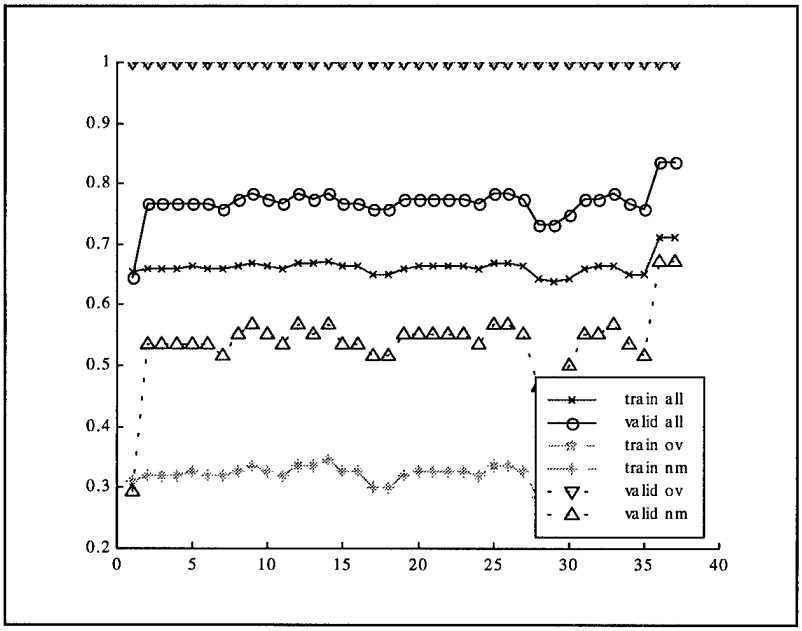

Figure 5-7. CA with Feature Reduction by Coefficient.

In all six models, similar patterns in CA were observed. First, all models suggested significant reduction of variables. In all six models significant drops in the

training set CA were seen with only one to five features remaining in the model. Next, on average, the validation set CA was much better than chance with many values in the 80% range. The maximum values were between 75% and almost 90%, in which almost 100% of the errors were made as the misclassification of nominal as overload. In all cases, the classification was highly skewed toward correct classification of overload. Finally, as evident in Table 5-9, feature saliency was consistent through all six models, with μbeta and some beta consistently in the top 5.

Table 5-9. Two-Class Linear Feature Saliency.

Linear Saliency by Discriminant Loadings and Normalized Coefficients						
	Model 1		Model 2		Model 3	
Rank	loading	coefficient	loading	coefficient	loading	coefficient
1	T7-ub	T7-ub	O2-ub	T7-ub	PZ-ub	T7-ub
2	T8-b	O2-ub	PZ-ub	O2-ub	FZ-ub	O2-ub
3	T8-ub	PZ-ub	T7-ub	PZ-ub	T7-ub	FZ-ub
4	F7-ub	T8-ub	T8-ub	T8-ub	O2-ub	F7-ub
5	F7-b	T8-b	T8-b	F7-ub	T8-b	F7-b
6	T7-b	O2-b	F7-ub	PZ-t	T8-ub	T7-b
7	O2-ub	IBLI	T7-b	F7-t	T7-b	T7-d
8	PZ-ub	F7-b	F7-b	T7-b	F7-ub	T8-ub
9	FZ-b	F7-a	FZ-b	F7-b	O2-b	brths
10	IBLI	PZ-b	O2-b	T7-t	FZ-b	PZ-ub

5.2.3 MLP ANN Modeling. The next step to be performed before ANNs could be trained included the selection of an optimal set of group salient features for this two-class modeling effort. To determine an optimal set, the two-class rankings of features by linear saliency presented above were used in conjunction with the *a priori* knowledge of the features identified as salient in the three-class ANN modeling efforts. For this research, a decision was made to try sets of input features with 2, 5, and 10 features. As identified in Figure 5-8, the top 2 features were μbeta at T7 and O2, and the top 5 features added

μbeta at T8 and PZ and eye-blink intervals. The top 10 features then added the number of eye blinks, μbeta at F7, and beta from F7, T7, and O2.

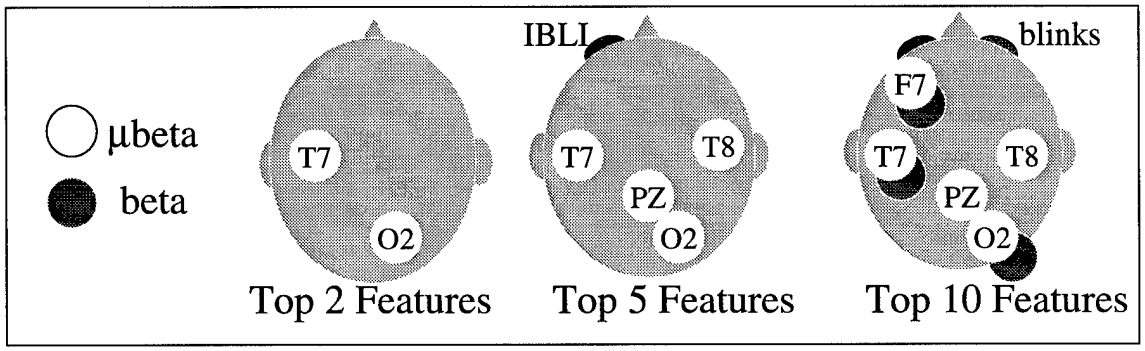

Figure 5-8. 2-Class Salient Group Features.

For each of the three models, ANNs were trained 30 times using sets of 2, 5, and 10 input features. ANN training was performed similarly to that described in Chapter 4, when multiple runs were completed. The primary differences include the change from three to two output nodes for the two-classes and the use of four of the subjects to train the ANN, with the remaining four subjects used only as validation. The average validation set classification accuracy by subject is included in Figure 5-9. Because all three selections of input features provided similar results, the CA presented is an average of the three modeling efforts. Subjects are ordered by CA, with subject 13 remaining as an outlier that appears to classify no better than chance.

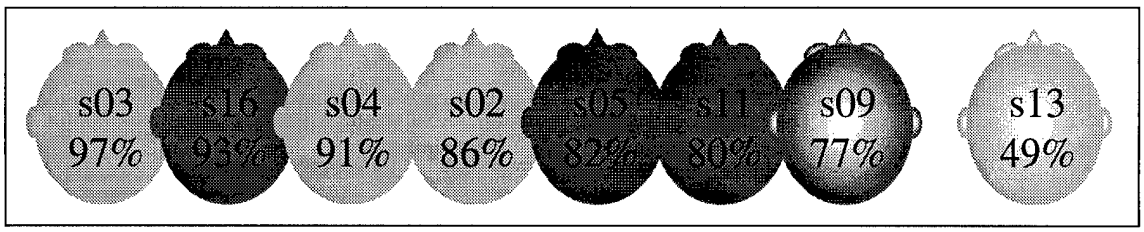

Figure 5-9. "One Net" Validation CA by Subject.

In addition, Table 5-10 provides the CA by subject for each of the models using different input features.

Table 5-10. "One Net" Validation CA

	Validation CA			
	2-f	5-f	10-f	mean
s03	95.9%	98.5%	96.9%	97.1%
s16	95.0%	95.3%	87.9%	92.7%
s04	91.2%	92.0%	89.0%	90.7%
s02	86.9%	84.8%	86.6%	86.1%
s05	90.9%	76.3%	80.1%	82.4%
s11	80.6%	79.2%	79.6%	79.8%
s09*	73.7%	79.6%	77.1%	76.8%
s13*	45.1%	50.6%	52.0%	49.2%
mean**	87.7%	86.5%	85.3%	**86.5%**
std dev**	8.1%	8.8%	6.9%	7.4%
*Average of 3 models **Does not include subject 13				

In addition to presenting the overall validation CA, the overload CA obtained when validating the ANNs is also presented. As was done for the overall validation CA, a figure is presented in which all subjects are ordered by Overload validation CA. Here, subject 13 remains a low outlier, while subject 09 appears to be in the middle of the pack.

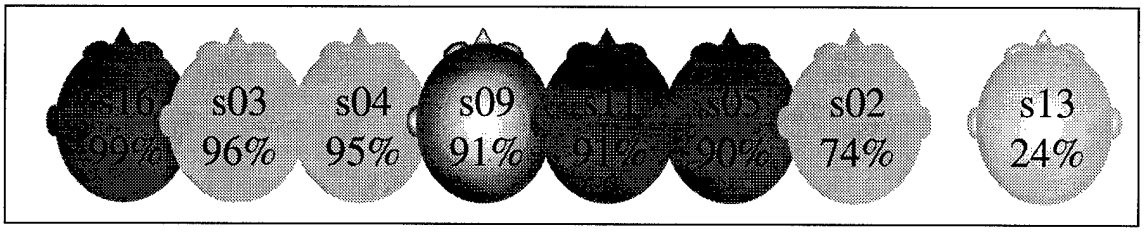

Figure 5-10. Overload "One Net" CA by Subject.

Finally, Table 5-11 provides the Overload CA by subject for each of the models using the different input feature sets.

Table 5-11. "One Net" Validation Overload CA

	Validation Set Overload CA			
	2-f	5-f	10-f	mean
s16	100.0%	98.5%	98.6%	99.0%
s03	95.5%	97.0%	95.3%	95.9%
s04	100.0%	94.8%	90.3%	95.0%
s09*	85.9%	95.5%	91.9%	91.1%
s11	92.1%	90.6%	89.9%	90.9%
s05	95.6%	85.0%	90.4%	90.3%
s02	77.5%	70.1%	75.0%	74.2%
s13*	19.0%	25.5%	25.9%	23.5%
mean**	92.4%	90.2%	90.2%	**90.9%**
std dev**	8.2%	10.0%	7.4%	8.0%
*Average of 3 models **Does not include subject 13				

5.2.4 *"Can One Net Fit All?"*. For this modeling experiment designed to detect overload after a nominal state, initial feasibility has been demonstrated for the potential use of one ANN to fit multiple subjects. Specifically, when training on four individuals and validating on four different individuals, promise is indicated for this two-group classification problem. This promise is demonstrated by an overall validation CA of 87% with an associated overload validation CA of over 90%, when excluding subject 13 as an apparent outlier.

VI. Conclusions and Recommendations

This final chapter summarizes the findings and conclusions of this research effort. The primary conclusions are related to the comparison of modeling techniques, selection of a salient set of group features, and assessing the feasibility of using one ANN to fit multiple test subjects. Topics of continued research are also presented.

6.1 Comparison of Models

In general, MLP ANNs were identified as superior classifiers in both the three-group and two-group classification efforts. Specifically, for the three-group classification problems, the MLP ANN was able to correctly classify medium workload much better than chance, while the discriminant models rarely classified any medium workload observations correctly at all. On the other hand, the discriminant models performed well for the two-group classification effort. In these efforts close to 100% accuracy was obtained for the overload level. Yet overall, these models performed with an average CA around 80% which was slightly lower than the average of about 90% obtained using ANNs.

Specific reasons for the performance differences of the two models stem from the assumptions required, the method used to calculate the parameters of the models, and the available form the model can take. In the case of discriminant analysis, the model assumes that the data is distributed as a multivariate Gaussian distribution. The model coefficients are then deterministically calculated using an estimate of the group means, covariance structure, and prior probability of classification. Thus, misclassification of the

medium workload was likely to occur when using the discriminant model, as the mean values for the various features were typically very close to either the low or overload means, and the covariance structure of the medium workload was slightly larger. Similarly, in the two-group classification effort, misclassifications were almost exclusively observed as a nominal condition being misclassified as overload. This was likely to be the result of a larger variance contained within the covariance structure as both low and medium observations were combined to form the nominal class.

Unlike the discriminant analysis, the MLP ANN models used for this effort make no assumptions about the structure of the input data. The ANNs adjust their parameters including weights and biases through supervised training. During the training, ANNs can adaptively learn thresholds and have the ability to generalize. This is similar to the ability of a Taylor series expansion to approximate any well-behaved function. Finally, MLP ANNs were able to classify all three workload groups even though they were presented with non-stationary psychophysiological data.

6.2 Feature Selection

6.2.1 Discriminant Feature Selection. Two heuristic algorithms were developed and utilized in this research effort. Both efforts used a similar methodology with one using discriminant loadings as a measure of feature saliency and the other using the coefficients obtained from normalized input variables as a measure of saliency. Both methods were effective at identifying those features that are linearly salient. In addition, both methods were consistent in selecting the same groups of features.

6.2.2 MLP ANN Feature Selection. Both linear and non-linear MLP ANN models were used to determine a salient set of features for use by an MLP ANN model. By including a linear modeling effort before each ANN modeling effort, this research has shown that while the linear models may be poor classifiers, they are useful for identifying linearly salient features. These linearly salient features then provide a good starting point for the input of a non-linear ANN model. The use of SNR feature selection was demonstrated to be robust across three hidden node architectures. The SNR feature selection also provided a set of top features that included those features identified by the linear saliency measures. Overall, the μbeta EEG features appeared most salient as a group, followed by eye-blink features. Beta features and respiration measures also appeared salient and consistently within a list of the top 10 most salient features.

6.3 "Can One Net Fit All?"

To determine if "one net could fit all" an experiment was devised using the available data to create a two-class problem aimed at detecting "overload" after "nominal" workload. In this experiment, four of the subjects were used to train a model, with the remaining four used to obtain an estimate of validation classification accuracy. Results were promising with a validation set CA of 87% including an overload CA of 92%, for seven of the eight subjects. Thus, the future use of "one net to fit all" appears to be feasible and may warrant additional research using other data sets.

6.4 *Workload Classification Findings*

This section includes a description of apparent temporal effects that may have a high level of influence on the ability to correctly classify an observation. Additionally, three methods were attempted in this research with the goal of increasing CA, and are described in the following section.

6.4.1 Temporal Effects. While the power derived from EEG is known to be non-stationary, the presentation order of workload appears to specifically have a significant effect on the values observed for EEG μbeta signals. Multiple plots of EEG were analyzed with emphasis on the salient μbeta features. In addition to an apparent time trend being present, inconsistencies in mean values could be correlated to the workload presentation order. This phenomenon definitely warrants further investigation. Further insights will allow for more efficient and possibly, more effective, experiments to be planned and analyzed in future research.

6.4.2 Methods to Improve CA. A quick feasibility study using one subject was performed to assess methods to improve CA. Three methods were used in this attempt. First, the input data was preprocessed into 20-second windows in the attempt to reduce the variance of the original features that were preprocessed using 10-second windows. Next, the possible classes were reduced from three to two classes. Low and medium workload were combined as one "nominal" class in which a test subject could complete all tasks. Finally, the two methods were combined where classification accuracy was found for a two-class problem using input features comprised of 20-second averages. Results of this study found the best improvement was found when reducing the number

of classes. Alone, the use of 20-second averages only slightly improved the three-class CA. When combined with the reduced classes, no significant improvement was obtained.

6.5 Recommendations for Future Efforts

While the data obtained for this research was time ordered, the models implemented did not make use of any temporal information. In fact, data preprocessing included the normalization of data every 15-minutes in the hope of removing significant temporal trends. Therefore, recommendations are made to incorporate the temporal information. In addition, other various research efforts could be accomplished using this data set in conjunction with the performance data of the test subjects or the similar set of data obtained during a second day of MAT-B testing.

6.5.1 Use Temporal Information. As has been mentioned, definite trends appear temporally within the data. Rather than trying to remove these effects, analysis methods should be employed that are able to use any temporal information. Two suggestions of models to try are as follows:

- Linear control chart or autoregressive linear model that monitors a score for variations in output
- Elman Recurrent or other ANN models with feedback

6.5.2 Feature Reduction Techniques. While feature reduction was performed through the use of both linear and ANN saliency screening, all feature reduction efforts were aimed at reducing the total number of psychophysyiological features used, rather than just reducing the dimensionality of the input to the classification models. Two methods are recommended for future research. The first suggestion is use of principal

component analysis (PCA) scores as model input. This could be performed using the 36 features processed for this effort, or the current 6 EEG locations could be augmented with up to 54 additional sites. Additionally, after looking at spatial plots of EEG power, a different approach may include just taking the PCA scores by frequency band or by location. The second method of feature reduction involves the use of factor analysis (FA) scores. FA scores are derived by rotating a chosen number of principal components, and may validate or suggest how to "smartly" formulate PCA scores. Like PCA, FA scores could be computed using any number of EEG locations as input, where the derived FA input features should be defendable by physiological theory.

6.5.3 Use of Additional MAT-B Data. Future efforts may include the analysis of MAT-B performance data. Specifically any correlation to the misclassification of an observation and the performance of an observed task could be interesting. Does any medium or low misclassification relate to incompleted tasks directly or with a time lag? Does any overload misclassification relate to periods of increased mission performance?

Additionally, with a similar scenario presented on the second day, a number of research efforts using the data from both days can be identified. The following list contains a few of these ideas.

- Are salient features consistent for an individual from day-to-day?
- Do day-to-day variations necessitate a uniquely trained ANN for an individual?
- Does the presentation of workload levels affect the ability to correctly classify workload for a given individual across two days?

6.5 Retrospect

During the course of this research effort, practical modeling experience was gained. Derived from these experiences are three useful insights for modeling and statistical pattern recognition. First, get familiar with the data. By viewing plots of all the raw data, the first μbeta observations were clearly detected as outliers. Also, these plots provided indications that μbeta and eye-blink features may be the most salient. More time could have also been spent in this phase, where time ordered plots of the data may have revealed inconsistent patterns early on. Second, while no model is 100% correct, some are useful. As evidence, the discriminant models were poor discriminators of the medium workload, yet provided useful insight for salient linear features. Third, it may be useful to develop less complicated models before trying more complicated ones. Lower order models may include fewer input variables, less complicated modeling techniques, or less classification states. Once a lower model is determined feasible, more complicated models can be attempted. As a minimum, creating lower order models will provide modeling experience with the data that may be useful for insight as to how the input features behave. This insight may also point toward the appropriate next model to attempt.

APPENDIX A: Validation Set Confusion Matrices by Subject

Table A-1. Subject 02 Confusion Matrices.

Subject 02

Confusion Matrix Key

low CA	low C'd as med	low C'd as OV	TRUE low
med C'd as low	**medium CA**	med C'd as OV	TRUE medium
OV C'd as low	OV C'd as med	**overload CA**	TRUE overload
Classified low	Classified medium	Classified overload	**Overall CA**

Individually Trained ANN **Internal Validation** individual features	2655 93.2%	130 4.6%	63 2.2%	2848					
	163 5.6%	2424 83.2%	328 11.3%	2915					
	36 1.2%	344 11.7%	2557 87.1%	2937					
	2854	2898	2948	8700 87.77%					
Individually Trained ANN **External Validation** individual features	4652 80.2%	721 12.4%	427 7.4%	5800					
	2799 48.3%	362 6.2%	2639 45.5%	5800					
	68 1.2%	2662 45.9%	3070 52.9%	5800					
	7519	3745	6136	17400 46.46%					
					Group Trained ANNs				
Individually Trained ANN **Internal Validation** 2-features	712 79.7%	139 15.6%	42 4.7%	893	Group Trained ANN **Internal Validation** 2-features	376 42.6%	304 34.4%	203 23.0%	883
	199 22.1%	341 37.9%	359 39.9%	899		250 28.8%	467 53.7%	152 17.5%	869
	35 4.3%	176 21.5%	607 74.2%	818		123 14.3%	391 45.6%	344 40.1%	858
	946	656	1008	2610 63.60%		749	1162	699	2610 45.48%
Individually Trained ANN **External Validation** 2-features	1371 78.8%	350 20.1%	19 1.1%	1740	Group Trained ANN **External Validation** 2-features	614 35.3%	681 39.1%	445 25.6%	1740
	649 37.3%	613 35.2%	478 27.5%	1740		365 21.0%	893 51.3%	482 27.7%	1740
	107 6.1%	332 19.1%	1301 74.8%	1740		282 16.2%	648 37.2%	810 46.6%	1740
	2127	1295	1798	5220 62.93%		1261	2222	1737	5220 44.39%
Individually Trained ANN **Internal Validation** 5-features	729 86.5%	87 10.3%	27 3.2%	843	Group Trained ANN **Internal Validation** 5-features	416 47.6%	254 29.1%	204 23.3%	874
	153 17.1%	543 60.6%	200 22.3%	896		199 22.8%	579 66.4%	94 10.8%	872
	35 4.0%	167 19.2%	669 76.8%	871		97 11.2%	388 44.9%	379 43.9%	864
	917	797	896	2610 74.37%		712	1221	677	2610 52.64%
Individually Trained ANN **External Validation** 5-features	823 47.3%	890 51.1%	27 1.6%	1740	Group Trained ANN **External Validation** 5-features	629 36.1%	999 57.4%	112 6.4%	1740
	387 22.2%	1017 58.4%	336 19.3%	1740		404 23.2%	1273 73.2%	63 3.6%	1740
	116 6.7%	302 17.4%	1322 76.0%	1740		498 28.6%	142 8.2%	1100 63.2%	1740
	1326	2209	1685	5220 60.57%		1531	2414	1275	5220 57.51%
Individually Trained ANN **Internal Validation** 10-features	786 91.7%	37 4.3%	34 4.0%	857	Group Trained ANN **Internal Vaidation** 10-features	632 73.5%	135 15.7%	93 10.8%	860
	87 10.2%	611 71.6%	155 18.2%	853		223 25.8%	572 66.3%	68 7.9%	863
	29 3.2%	151 16.8%	720 80.0%	900		110 12.4%	367 41.4%	410 46.2%	887
	902	799	909	2610 81.11%		965	1074	571	2610 61.84%

Table A-1 Continued. Subject 02 Confusion Matrices.

Individually Trained ANN **Internal Validation** 10-features	786 91.7%	37 4.3%	34 4.0%	857	Group Trained ANN **Internal Vaidation** 10-features	632 73.5%	135 15.7%	93 10.8%	860
	87 10.2%	611 71.6%	155 18.2%	853		223 25.8%	572 66.3%	68 7.9%	863
	29 3.2%	151 16.8%	720 80.0%	900		110 12.4%	367 41.4%	410 46.2%	887
	902	799	909	2610 81.11%		965	1074	571	2610 61.84%
Individually Trained ANN **External Validation** 10-features	1259 72.4%	379 21.8%	102 5.9%	1740	Group Trained ANN **External Validation** 10-features	844 48.5%	779 44.8%	117 6.7%	1740
	209 12.0%	1052 60.5%	479 27.5%	1740		410 23.6%	1201 69.0%	129 7.4%	1740
	57 3.3%	233 13.4%	1450 83.3%	1740		200 11.5%	210 12.1%	1330 76.4%	1740
	1525	1664	2031	5220 72.05%		1454	2190	1576	5220 64.66%
Individually Trained ANN **Internal Validation** 15-features	785 93.1%	37 4.4%	21 2.5%	843	Group Trained ANN **Internal Validation** 15-features	710 85.9%	62 7.5%	55 6.7%	827
	31 3.5%	765 85.9%	95 10.7%	891		210 23.6%	632 71.1%	47 5.3%	889
	13 1.5%	88 10.0%	775 88.5%	876		74 8.3%	246 27.5%	574 64.2%	894
	829	890	891	2610 89.08%		994	940	676	2610 73.41%
Individually Trained ANN **External Validation** 15-features	1037 59.6%	633 36.4%	70 4.0%	1740	Group Trained ANN **External Validation** 15-features	930 53.4%	760 43.7%	50 2.9%	1740
	602 34.6%	310 17.8%	828 47.6%	1740		457 26.3%	756 43.4%	527 30.3%	1740
	66 3.8%	399 22.9%	1275 73.3%	1740		252 14.5%	384 22.1%	1104 63.4%	1740
	1705	1342	2173	5220 50.23%		1639	1900	1681	5220 53.45%

Table A-2. Subject 03 Confusion Matrices.

Individually Trained ANN **Internal Validation** individual features	2393 81.4%	521 17.7%	27 0.9%	2941
	600 20.6%	2094 71.8%	221 7.6%	2915
	9 0.3%	133 4.7%	2702 95.0%	2844
	3002	2748	2950	8700 82.63%
Individually Trained ANN **External Validation** individual features	256 4.4%	5427 93.6%	117 2.0%	5800
	4532 78.1%	985 17.0%	283 4.9%	5800
	8 0.1%	812 14.0%	4980 85.9%	5800
	4796	7224	5380	17400 35.75%

Subject 03

Confusion Matrix Key

low CA	low C'd as med	low C'd as OV	TRUE low
med C'd as low	**medium CA**	med C'd as OV	TRUE medium
OV C'd as low	OV C'd as med	**overload CA**	TRUE overload
Classified low	Classified medium	Classified overload	**Overall CA**

Table A-2 Continued. Subject 03 Confusion Matrices.

Individually Trained ANN **Internal Validation** 2-features	626	237	0	863	Group Trained ANN **Internal Validation** 2-features	535	306	5	846
	72.5%	27.5%	0.0%			63.2%	36.2%	0.6%	
	201	601	39	841		230	572	52	854
	23.9%	71.5%	4.6%			26.9%	67.0%	6.1%	
	1	64	841	906		45	67	798	910
	0.1%	7.1%	92.8%			4.9%	7.4%	87.7%	
	828	902	880	2610		810	945	855	2610
				79.23%					72.99%
Individually Trained ANN **External Validation** 2-features	297	1415	28	1740	Group Trained ANN **External Validation** 2-features	837	876	27	1740
	17.1%	81.3%	1.6%			48.1%	50.3%	1.6%	
	906	787	47	1740		1066	552	122	1740
	52.1%	45.2%	2.7%			61.3%	31.7%	7.0%	
	0	222	1518	1740		30	102	1608	1740
	0.0%	12.8%	87.2%			1.7%	5.9%	92.4%	
	1203	2424	1593	5220		1933	1530	1757	5220
				49.85%					57.41%
Individually Trained ANN **Internal Validation** 5-features	678	209	2	889	Group Trained ANN **Internal Validation** 5-features	555	347	0	902
	76.3%	23.5%	0.2%			61.5%	38.5%	0.0%	
	201	606	44	851		222	549	96	867
	23.6%	71.2%	5.2%			25.6%	63.3%	11.1%	
	3	41	826	870		28	60	753	841
	0.3%	4.7%	94.9%			3.3%	7.1%	89.5%	
	882	856	872	2610		805	956	849	2610
				80.84%					71.15%
Individually Trained ANN **External Validation** 5-features	384	1295	61	1740	Group Trained ANN **External Validation** 5-features	645	1055	40	1740
	22.1%	74.4%	3.5%			37.1%	60.6%	2.3%	
	1217	501	22	1740		1032	676	32	1740
	69.9%	28.8%	1.3%			59.3%	38.9%	1.8%	
	0	261	1479	1740		13	174	1553	1740
	0.0%	15.0%	85.0%			0.7%	10.0%	89.3%	
	1601	2057	1562	5220		1690	1905	1625	5220
				45.29%					55.06%
Individually Trained ANN **Internal Validation** 10-features	681	157	0	838	Group Trained ANN **Internal Vaidation** 10-features	610	302	7	919
	81.3%	18.7%	0.0%			66.4%	32.9%	0.8%	
	179	646	70	895		200	576	69	845
	20.0%	72.2%	7.8%			23.7%	68.2%	8.2%	
	2	60	815	877		20	94	732	846
	0.2%	6.8%	92.9%			2.4%	11.1%	86.5%	
	862	863	885	2610		830	972	808	2610
				82.07%					73.49%
Individually Trained ANN **External Validation** 10-features	147	1527	66	1740	Group Trained ANN **External Validation** 10-features	434	1250	56	1740
	8.4%	87.8%	3.8%			24.9%	71.8%	3.2%	
	1468	244	28	1740		1065	601	74	1740
	84.4%	14.0%	1.6%			61.2%	34.5%	4.3%	
	17	261	1462	1740		5	269	1466	1740
	1.0%	15.0%	84.0%			0.3%	15.5%	84.3%	
	1632	2032	1556	5220		1504	2120	1596	5220
				35.50%					47.91%
Individually Trained ANN **Internal Validation** 15-features	680	156	4	840	Group Trained ANN **Internal Validation** 15-features	585	274	24	883
	81.0%	18.6%	0.5%			66.3%	31.0%	2.7%	
	162	627	70	859		209	574	104	887
	18.9%	73.0%	8.1%			23.6%	64.7%	11.7%	
	3	72	836	911		41	90	709	840
	0.3%	7.9%	91.8%			4.9%	10.7%	84.4%	
	845	855	910	2610		835	938	837	2610
				82.11%					71.57%
Individually Trained ANN **External Validation** 15-features	60	1598	82	1740	Group Trained ANN **External Validation** 15-features	650	1045	45	1740
	3.4%	91.8%	4.7%			37.4%	60.1%	2.6%	
	1441	288	11	1740		886	748	106	1740
	82.8%	16.6%	0.6%			50.9%	43.0%	6.1%	
	4	257	1479	1740		5	160	1575	1740
	0.2%	14.8%	85.0%			0.3%	9.2%	90.5%	
	1505	2143	1572	5220		1541	1953	1726	5220
				35.00%					56.95%

Table A-3. Subject 04 Confusion Matrices.

Individually Trained ANN	2551	310	41	2902	**Subject 04**				
	87.9%	10.7%	1.4%						
Internal Validation	228	2116	534	2878					
	7.9%	73.5%	18.6%		Confusion Matrix Key				
individual features	40	340	2540	2920	**low CA**	low C'd as med	low C'd as OV	TRUE low	
	1.4%	11.6%	87.0%						
	2819	2766	3115	8700	med C'd as low	**medium CA**	med C'd as OV	TRUE medium	
				82.84%					
Individually Trained ANN	5047	504	249	5800	OV C'd as low	OV C'd as med	**overload CA**	TRUE overload	
	87.0%	8.7%	4.3%						
External Validation	1569	3475	756	5800	Classified low	Classified medium	Classified overload	**Overall CA**	
	27.1%	59.9%	13.0%						
individual features	4	432	5364	5800					
	0.1%	7.4%	92.5%						
	6620	4411	6369	17400					
				79.80%					
Individually Trained ANN	503	257	87	847	Group Trained ANN	561	142	142	845
	59.4%	30.3%	10.3%			66.4%	16.8%	16.8%	
Internal Validation	81	550	260	891	**Internal Validation**	144	446	317	907
	9.1%	61.7%	29.2%			15.9%	49.2%	35.0%	
2-features	19	268	585	872	2-features	14	172	672	858
	2.2%	30.7%	67.1%			1.6%	20.0%	78.3%	
	603	1075	932	2610		719	760	1131	2610
				62.76%					64.33%
Individually Trained ANN	1033	566	141	1740	Group Trained ANN	1192	452	96	1740
	59.4%	32.5%	8.1%			68.5%	26.0%	5.5%	
External Validation	267	900	573	1740	**External Validation**	514	640	586	1740
	15.3%	51.7%	32.9%			29.5%	36.8%	33.7%	
2-features	50	576	1114	1740	2-features	76	181	1483	1740
	2.9%	33.1%	64.0%			4.4%	10.4%	85.2%	
	1350	2042	1828	5220		1782	1273	2165	5220
				58.37%					63.51%
Individually Trained ANN	675	161	34	870	Group Trained ANN	587	234	44	865
	77.6%	18.5%	3.9%			67.9%	27.1%	5.1%	
Internal Validation	90	536	268	894	**Internal Validation**	110	426	319	855
	10.1%	60.0%	30.0%			12.9%	49.8%	37.3%	
5-features	1	151	694	846	5-features	6	130	754	890
	0.1%	17.8%	82.0%			0.7%	14.6%	84.7%	
	766	848	996	2610		703	790	1117	2610
				72.99%					67.70%
Individually Trained ANN	1249	381	110	1740	Group Trained ANN	1118	504	118	1740
	71.8%	21.9%	6.3%			64.3%	29.0%	6.8%	
External Validation	284	1015	441	1740	**External Validation**	468	802	470	1740
	16.3%	58.3%	25.3%			26.9%	46.1%	27.0%	
5-features	18	234	1488	1740	5-features	54	87	1599	1740
	1.0%	13.4%	85.5%			3.1%	5.0%	91.9%	
	1551	1630	2039	5220		1640	1393	2187	5220
				71.88%					67.41%

Table A-3 Continued. Subject 04 Confusion Matrices.

Individually Trained ANN **Internal Validation** 10-features	728 81.6%	158 17.7%	6 0.7%	892	Group Trained ANN **Internal Vaidation** 10-features	632 71.4%	210 23.7%	43 4.9%	885
	67 7.6%	620 70.2%	196 22.2%	883		112 13.1%	501 58.4%	245 28.6%	858
	4 0.5%	114 13.7%	717 85.9%	835		10 1.2%	102 11.8%	755 87.1%	867
	799	892	919	2610 79.12%		754	813	1043	2610 72.34%
Individually Trained ANN **External Validation** 10-features	1256 72.2%	359 20.6%	125 7.2%	1740	Group Trained ANN **External Validation** 10-features	1244 71.5%	350 20.1%	146 8.4%	1740
	376 21.6%	1096 63.0%	268 15.4%	1740		471 27.1%	882 50.7%	387 22.2%	1740
	6 0.3%	143 8.2%	1591 91.4%	1740		89 5.1%	85 4.9%	1566 90.0%	1740
	1638	1598	1984	5220 75.54%		1804	1317	2099	5220 70.73%
Individually Trained ANN **Internal Validation** 15-features	751 86.9%	104 12.0%	9 1.0%	864	Group Trained ANN **Internal Validation** 15-features	724 76.6%	203 21.5%	18 1.9%	945
	57 6.4%	670 75.7%	158 17.9%	885		92 11.3%	496 61.1%	224 27.6%	812
	5 0.6%	108 12.5%	748 86.9%	861		3 0.4%	97 11.4%	753 88.3%	853
	813	882	915	2610 83.10%		819	796	995	2610 75.59%
Individually Trained ANN **External Validation** 15-features	1365 78.4%	249 14.3%	126 7.2%	1740	Group Trained ANN **External Validation** 15-features	1280 73.6%	326 18.7%	134 7.7%	1740
	444 25.5%	1021 58.7%	275 15.8%	1740		520 29.9%	882 50.7%	338 19.4%	1740
	1 0.1%	145 8.3%	1594 91.6%	1740		16 0.9%	135 7.8%	1589 91.3%	1740
	1810	1415	1995	5220 76.25%		1816	1343	2061	5220 71.86%

Table A-4. Subject 05 Confusion Matrices.

Individually Trained ANN **Internal Validation** individual features	2424 81.8%	509 17.2%	32 1.1%	2965
	379 13.3%	1914 67.3%	551 19.4%	2844
	3 0.1%	459 15.9%	2429 84.0%	2891
	2806	2882	3012	8700 77.78%
Individually Trained ANN **External Validation** individual features	5318 91.7%	482 8.3%	0 0.0%	5800
	878 15.1%	4567 78.7%	355 6.1%	5800
	13 0.2%	517 8.9%	5270 90.9%	5800
	6209	5566	5625	17400 87.10%

Subject 05

Confusion Matrix Key

low CA	low C'd as med	low C'd as OV	TRUE low
med C'd as low	**medium CA**	med C'd as OV	TRUE medium
OV C'd as low	OV C'd as med	**overload CA**	TRUE overload
Classified low	Classified medium	Classified overload	**Overall CA**

Table A-4 Continued. Subject 05 Confusion Matrices.

Individually Trained ANN **Internal Validation** 2-features	690	87	111	888	Group Trained ANN **Internal Validation** 2-features	652	132	64	848
	77.7%	9.8%	12.5%			76.9%	15.6%	7.5%	
	148	289	427	864		140	484	283	907
	17.1%	33.4%	49.4%			15.4%	53.4%	31.2%	
	25	130	703	858		30	296	529	855
	2.9%	15.2%	81.9%			3.5%	34.6%	61.9%	
	863	506	1241	2610		822	912	876	2610
				64.44%					63.79%
Individually Trained ANN **External Validation** 2-features	1101	227	412	1740	Group Trained ANN **External Validation** 2-features	1096	443	201	1740
	63.3%	13.0%	23.7%			63.0%	25.5%	11.6%	
	1129	337	274	1740		969	742	29	1740
	64.9%	19.4%	15.7%			55.7%	42.6%	1.7%	
	0	140	1600	1740		14	117	1609	1740
	0.0%	8.0%	92.0%			0.8%	6.7%	92.5%	
	2230	704	2286	5220		2079	1302	1839	5220
				58.20%					66.03%
Individually Trained ANN **Internal Validation** 5-features	725	97	26	848	Group Trained ANN **Internal Validation** 5-features	664	117	58	839
	85.5%	11.4%	3.1%			79.1%	13.9%	6.9%	
	91	621	129	841		124	491	267	882
	10.8%	73.8%	15.3%			14.1%	55.7%	30.3%	
	4	96	821	921		16	291	582	889
	0.4%	10.4%	89.1%			1.8%	32.7%	65.5%	
	820	814	976	2610		804	899	907	2610
				83.03%					66.55%
Individually Trained ANN **External Validation** 5-features	1509	228	3	1740	Group Trained ANN **External Validation** 5-features	1201	426	113	1740
	86.7%	13.1%	0.2%			69.0%	24.5%	6.5%	
	384	1160	196	1740		661	1075	4	1740
	22.1%	66.7%	11.3%			38.0%	61.8%	0.2%	
	1	49	1690	1740		1	77	1662	1740
	0.1%	2.8%	97.1%			0.1%	4.4%	95.5%	
	1894	1437	1889	5220		1863	1578	1779	5220
				83.51%					75.44%
Individually Trained ANN **Internal Validation** 10-features	737	109	12	858	Group Trained ANN **Internal Vaidation** 10-features	696	134	39	869
	85.9%	12.7%	1.4%			80.1%	15.4%	4.5%	
	71	652	155	878		126	490	265	881
	8.1%	74.3%	17.7%			14.3%	55.6%	30.1%	
	0	96	778	874		33	234	593	860
	0.0%	11.0%	89.0%			3.8%	27.2%	69.0%	
	808	857	945	2610		855	858	897	2610
				83.03%					68.16%
Individually Trained ANN **External Validation** 10-features	1443	297	0	1740	Group Trained ANN **External Validation** 10-features	1127	555	58	1740
	82.9%	17.1%	0.0%			64.8%	31.9%	3.3%	
	503	868	369	1740		800	898	42	1740
	28.9%	49.9%	21.2%			46.0%	51.6%	2.4%	
	4	253	1483	1740		20	48	1672	1740
	0.2%	14.5%	85.2%			1.1%	2.8%	96.1%	
	1950	1418	1852	5220		1947	1501	1772	5220
				72.68%					70.82%
Individually Trained ANN **Internal Validation** 15-features	740	99	24	863	Group Trained ANN **Internal Validation** 15-features	640	172	49	861
	85.7%	11.5%	2.8%			74.3%	20.0%	5.7%	
	88	614	169	871		103	465	311	879
	10.1%	70.5%	19.4%			11.7%	52.9%	35.4%	
	2	103	771	876		31	193	646	870
	0.2%	11.8%	88.0%			3.6%	22.2%	74.3%	
	830	816	964	2610		774	830	1006	2610
				81.42%					67.09%
Individually Trained ANN **External Validation** 15-features	1354	385	1	1740	Group Trained ANN **External Validation** 15-features	1058	625	57	1740
	77.8%	22.1%	0.1%			60.8%	35.9%	3.3%	
	556	840	344	1740		744	921	75	1740
	32.0%	48.3%	19.8%			42.8%	52.9%	4.3%	
	6	372	1362	1740		13	69	1658	1740
	0.3%	21.4%	78.3%			0.7%	4.0%	95.3%	
	1916	1597	1707	5220		1815	1615	1790	5220
				68.12%					69.67%

Table A-5. Subject 09 Confusion Matrices.

Individually Trained ANN	2814	75	2	2891	**Subject 09**				
	97.3%	2.6%	0.1%						
Internal Validation	65	2852	7	2924					
	2.2%	97.5%	0.2%		Confusion Matrix Key				
individual features	0	2	2883	2885	**low CA**	low C'd as med	low C'd as OV	TRUE low	
	0.0%	0.1%	99.9%		med C'd as low	**medium CA**	med C'd as OV	TRUE medium	
	2879	2929	2892	8700	OV C'd as low	OV C'd as med	**overload CA**	TRUE overload	
				98.26%	Classified low	Classified medium	Classified overload	**Overall CA**	
Individually Trained ANN	4671	1129	0	5800					
	80.5%	19.5%	0.0%						
External Validation	1174	4615	11	5800					
	20.2%	79.6%	0.2%						
individual features	0	191	5609	5800					
	0.0%	3.3%	96.7%						
	5845	5935	5620	17400					
				85.60%					
Individually Trained ANN	752	112	11	875	Group Trained ANN	792	68	0	860
	85.9%	12.8%	1.3%			92.1%	7.9%	0.0%	
Internal Validation	74	694	100	868	**Internal Validation**	92	592	175	859
	8.5%	80.0%	11.5%			10.7%	68.9%	20.4%	
2-features	0	75	792	867	2-features	14	41	836	891
	0.0%	8.7%	91.3%			1.6%	4.6%	93.8%	
	826	881	903	2610		898	701	1011	2610
				85.75%					85.06%
Individually Trained ANN	1498	228	14	1740	Group Trained ANN	1571	169	0	1740
	86.1%	13.1%	0.8%			90.3%	9.7%	0.0%	
External Validation	182	1340	218	1740	**External Validation**	219	1099	422	1740
	10.5%	77.0%	12.5%			12.6%	63.2%	24.3%	
2-features	0	136	1604	1740	2-features	19	43	1678	1740
	0.0%	7.8%	92.2%			1.1%	2.5%	96.4%	
	1680	1704	1836	5220		1809	1311	2100	5220
				85.10%					83.30%
Individually Trained ANN	856	32	0	888	Group Trained ANN	797	57	1	855
	96.4%	3.6%	0.0%			93.2%	6.7%	0.1%	
Internal Validation	39	786	37	862	**Internal Validation**	53	631	198	882
	4.5%	91.2%	4.3%			6.0%	71.5%	22.4%	
5-features	5	26	829	860	5-features	10	35	828	873
	0.6%	3.0%	96.4%			1.1%	4.0%	94.8%	
	900	844	866	2610		860	723	1027	2610
				94.67%					86.44%
Individually Trained ANN	1411	329	0	1740	Group Trained ANN	1581	159	0	1740
	81.1%	18.9%	0.0%			90.9%	9.1%	0.0%	
External Validation	438	1096	206	1740	**External Validation**	227	1013	500	1740
	25.2%	63.0%	11.8%			13.0%	58.2%	28.7%	
5-features	0	241	1499	1740	5-features	6	26	1708	1740
	0.0%	13.9%	86.1%			0.3%	1.5%	98.2%	
	1849	1666	1705	5220		1814	1198	2208	5220
				76.74%					82.41%

Table A-5 Continued. Subject 09 Confusion Matrices.

Individually Trained ANN	859	21	10	890	Group Trained ANN	809	57	3	869
	96.5%	2.4%	1.1%			93.1%	6.6%	0.3%	
Internal Validation	40	794	4	838	**Internal Vaidation**	78	651	135	864
	4.8%	94.7%	0.5%			9.0%	75.3%	15.6%	
10-features	0	7	875	882	10-features	31	5	841	877
	0.0%	0.8%	99.2%			3.5%	0.6%	95.9%	
	899	822	889	2610		918	713	979	2610
				96.86%					88.16%
Individually Trained ANN	1432	306	2	1740	Group Trained ANN	1596	144	0	1740
	82.3%	17.6%	0.1%			91.7%	8.3%	0.0%	
External Validation	311	1429	0	1740	**External Validation**	164	1150	426	1740
	17.9%	82.1%	0.0%			9.4%	66.1%	24.5%	
10-features	2	101	1637	1740	10-features	1	2	1737	1740
	0.1%	5.8%	94.1%			0.1%	0.1%	99.8%	
	1745	1836	1639	5220		1761	1296	2163	5220
				86.17%					85.88%
Individually Trained ANN	831	25	0	856	Group Trained ANN	842	56	0	898
	97.1%	2.9%	0.0%			93.8%	6.2%	0.0%	
Internal Validation	38	847	3	888	**Internal Validation**	88	624	144	856
	4.3%	95.4%	0.3%			10.3%	72.9%	16.8%	
15-features	0	0	866	866	15-features	10	12	834	856
	0.0%	0.0%	100.0%			1.2%	1.4%	97.4%	
	869	872	869	2610		940	692	978	2610
				97.47%					88.12%
Individually Trained ANN	1367	373	0	1740	Group Trained ANN	1608	118	14	1740
	78.6%	21.4%	0.0%			92.4%	6.8%	0.8%	
External Validation	293	1447	0	1740	**External Validation**	156	1172	412	1740
	16.8%	83.2%	0.0%			9.0%	67.4%	23.7%	
15-features	0	57	1683	1740	15-features	5	7	1728	1740
	0.0%	3.3%	96.7%			0.3%	0.4%	99.3%	
	1660	1877	1683	5220		1769	1297	2154	5220
				86.15%					86.36%

Table A-6. Subject 11 Confusion Matrices.

Individually Trained ANN	2645	241	10	2896
	91.3%	8.3%	0.3%	
Internal Validation	304	2286	338	2928
	10.4%	78.1%	11.5%	
individual features	2	275	2599	2876
	0.1%	9.6%	90.4%	
	2951	2802	2947	8700
				86.55%
Individually Trained ANN	5482	298	20	5800
	94.5%	5.1%	0.3%	
External Validation	5	1254	4541	5800
	0.1%	21.6%	78.3%	
individual features	0	882	4918	5800
	0.0%	15.2%	84.8%	
	5487	2434	9479	17400
				66.98%

Subject 11

Confusion Matrix Key

low CA	low C'd as med	low C'd as OV	TRUE low
med C'd as low	**medium CA**	med C'd as OV	TRUE medium
OV C'd as low	OV C'd as med	**overload CA**	TRUE overload
Classified low	Classified medium	Classified overload	**Overall CA**

Table A-6 Continued. Subject 11 Confusion Matrices.

Individually Trained ANN **Internal Validation** 2-features	670 76.0%	185 21.0%	26 3.0%	881	Group Trained ANN **Internal Validation** 2-features	703 79.2%	163 18.4%	22 2.5%	888
	185 21.1%	573 65.3%	119 13.6%	877		265 30.7%	493 57.2%	104 12.1%	862
	31 3.6%	93 10.9%	728 85.4%	852		10 1.2%	104 12.1%	746 86.7%	860
	886	851	873	2610 75.52%		978	760	872	2610 74.41%
Individually Trained ANN **External Validation** 2-features	1597 91.8%	143 8.2%	0 0.0%	1740	Group Trained ANN **External Validation** 2-features	1659 95.3%	81 4.7%	0 0.0%	1740
	59 3.4%	802 46.1%	879 50.5%	1740		19 1.1%	858 49.3%	863 49.6%	1740
	65 3.7%	44 2.5%	1631 93.7%	1740		7 0.4%	89 5.1%	1644 94.5%	1740
	1721	989	2510	5220 77.20%		1685	1028	2507	5220 79.71%
Individually Trained ANN **Internal Validation** 5-features	706 82.0%	98 11.4%	57 6.6%	861	Group Trained ANN **Internal Validation** 5-features	698 78.3%	173 19.4%	20 2.2%	891
	200 22.4%	410 46.0%	282 31.6%	892		158 18.7%	567 67.1%	120 14.2%	845
	73 8.5%	261 30.5%	523 61.0%	857		1 0.1%	101 11.6%	772 88.3%	874
	979	769	862	2610 62.80%		857	841	912	2610 78.05%
Individually Trained ANN **External Validation** 5-features	633 36.4%	703 40.4%	404 23.2%	1740	Group Trained ANN **External Validation** 5-features	1655 95.1%	85 4.9%	0 0.0%	1740
	1046 60.1%	337 19.4%	357 20.5%	1740		7 0.4%	917 52.7%	816 46.9%	1740
	125 7.2%	892 51.3%	723 41.6%	1740		0 0.0%	21 1.2%	1719 98.8%	1740
	1804	1932	1484	5220 32.43%		1662	1023	2535	5220 82.20%
Individually Trained ANN **Internal Validation** 10-features	775 87.7%	102 11.5%	7 0.8%	884	Group Trained ANN **Internal Validation** 10-features	736 84.6%	122 14.0%	12 1.4%	870
	116 13.4%	661 76.5%	87 10.1%	864		242 27.8%	510 58.5%	120 13.8%	872
	0 0.0%	70 8.1%	792 91.9%	862		24 2.8%	40 4.6%	804 92.6%	868
	891	833	886	2610 85.36%		1002	672	936	2610 78.54%
Individually Trained ANN **External Validation** 10-features	1630 93.7%	105 6.0%	5 0.3%	1740	Group Trained ANN **External Validation** 10-features	1712 98.4%	27 1.6%	1 0.1%	1740
	0 0.0%	850 48.9%	890 51.1%	1740		107 6.1%	637 36.6%	996 57.2%	1740
	2 0.1%	91 5.2%	1647 94.7%	1740		10 0.6%	10 0.6%	1720 98.9%	1740
	1632	1046	2542	5220 79.06%		1829	674	2717	5220 77.95%
Individually Trained ANN **Internal Validation** 15-features	725 83.6%	129 14.9%	13 1.5%	867	Group Trained ANN **Internal Validation** 15-features	705 82.3%	136 15.9%	16 1.9%	857
	152 17.5%	654 75.2%	64 7.4%	870		252 28.6%	519 58.9%	110 12.5%	881
	1 0.1%	76 8.7%	796 91.2%	873		13 1.5%	62 7.1%	797 91.4%	872
	878	859	873	2610 83.33%		970	717	923	2610 77.43%
Individually Trained ANN **External Validation** 15-features	1632 93.8%	90 5.2%	18 1.0%	1740	Group Trained ANN **External Validation** 15-features	1651 94.9%	89 5.1%	0 0.0%	1740
	1 0.1%	1002 57.6%	737 42.4%	1740		67 3.9%	765 44.0%	908 52.2%	1740
	4 0.2%	140 8.0%	1596 91.7%	1740		5 0.3%	15 0.9%	1720 98.9%	1740
	1637	1232	2351	5220 81.03%		1723	869	2628	5220 79.23%

Table A-7. Subject 13 Confusion Matrices.

Individually Trained ANN **Internal Validation** individual features	2376 82.3%	415 14.4%	97 3.4%	2888					
	575 19.7%	1428 49.0%	913 31.3%	2916					
	19 0.7%	584 20.2%	2293 79.2%	2896					
	2970	2427	3303	8700 70.08%					
Individually Trained ANN **External Validation** individual features	2834 48.9%	2329 40.2%	637 11.0%	5800					
	3648 62.9%	1305 22.5%	847 14.6%	5800					
	201 3.5%	1475 25.4%	4124 71.1%	5800					
	6683	5109	5608	17400 47.49%					
Individually Trained ANN **Internal Validation** 2-features	637 73.7%	129 14.9%	98 11.3%	864	Group Trained ANN **Internal Validation** 2-features	198 22.6%	460 52.5%	218 24.9%	876
	248 27.6%	271 30.1%	381 42.3%	900		201 23.1%	515 59.3%	153 17.6%	869
	124 14.7%	300 35.5%	422 49.9%	846		176 20.3%	481 55.6%	208 24.0%	865
	1009	700	901	2610 50.96%		575	1456	579	2610 35.29%
Individually Trained ANN **External Validation** 2-features	872 50.1%	513 29.5%	355 20.4%	1740	Group Trained ANN **External Validation** 2-features	695 39.9%	647 37.2%	398 22.9%	1740
	1026 59.0%	445 25.6%	269 15.5%	1740		324 18.6%	846 48.6%	570 32.8%	1740
	54 3.1%	574 33.0%	1112 63.9%	1740		154 8.9%	1044 60.0%	542 31.1%	1740
	1952	1532	1736	5220 46.53%		1173	2537	1510	5220 39.90%
Individually Trained ANN **Internal Validation** 5-features	706 82.0%	98 11.4%	57 6.6%	861	Group Trained ANN **Internal Validation** 5-features	400 48.0%	239 28.7%	194 23.3%	833
	200 22.4%	410 46.0%	282 31.6%	892		237 26.5%	499 55.8%	158 17.7%	894
	73 8.5%	261 30.5%	523 61.0%	857		194 22.0%	489 55.4%	200 22.7%	883
	979	769	862	2610 62.80%		831	1227	552	2610 42.11%
Individually Trained ANN **External Validation** 5-features	633 36.4%	703 40.4%	404 23.2%	1740	Group Trained ANN **External Validation** 5-features	606 34.8%	755 43.4%	379 21.8%	1740
	1046 60.1%	337 19.4%	357 20.5%	1740		423 24.3%	640 36.8%	677 38.9%	1740
	125 7.2%	892 51.3%	723 41.6%	1740		182 10.5%	1148 66.0%	410 23.6%	1740
	1804	1932	1484	5220 32.43%		1211	2543	1466	5220 31.72%

Subject 13

Confusion Matrix Key

low CA	low C'd as med	low C'd as OV	TRUE low
med C'd as low	**medium CA**	med C'd as OV	TRUE medium
OV C'd as low	OV C'd as med	**overload CA**	TRUE overload
Classified low	Classified medium	Classified overload	**Overall CA**

Table A-7 Continued. Subject 13 Confusion Matrices.

Individually Trained ANN **Internal Validation** 10-features	710	96	44	850	Group Trained ANN **Internal Vaidation** 10-features	504	220	136	860
	83.5%	11.3%	5.2%			58.6%	25.6%	15.8%	
	163	471	240	874		223	530	173	926
	18.6%	53.9%	27.5%			24.1%	57.2%	18.7%	
	21	242	623	886		157	475	192	824
	2.4%	27.3%	70.3%			19.1%	57.6%	23.3%	
	894	809	907	2610		884	1225	501	2610
				69.12%					46.97%
Individually Trained ANN **External Validation** 10-features	934	367	439	1740	Group Trained ANN **External Validation** 10-features	913	669	158	1740
	53.7%	21.1%	25.2%			52.5%	38.4%	9.1%	
	1277	240	223	1740		576	664	500	1740
	73.4%	13.8%	12.8%			33.1%	38.2%	28.7%	
	73	694	973	1740		211	880	649	1740
	4.2%	39.9%	55.9%			12.1%	50.6%	37.3%	
	2284	1301	1635	5220		1700	2213	1307	5220
				41.13%					42.64%
Individually Trained ANN **Internal Validation** 15-features	754	83	23	860	Group Trained ANN **Internal Validation** 15-features	468	219	163	850
	87.7%	9.7%	2.7%			55.1%	25.8%	19.2%	
	147	488	235	870		142	557	154	853
	16.9%	56.1%	27.0%			16.6%	65.3%	18.1%	
	21	179	680	880		158	453	296	907
	2.4%	20.3%	77.3%			17.4%	49.9%	32.6%	
	922	750	938	2610		768	1229	613	2610
				73.64%					50.61%
Individually Trained ANN **External Validation** 15-features	837	532	371	1740	Group Trained ANN **External Validation** 15-features	1003	520	217	1740
	48.1%	30.6%	21.3%			57.6%	29.9%	12.5%	
	1172	356	212	1740		335	860	545	1740
	67.4%	20.5%	12.2%			19.3%	49.4%	31.3%	
	102	441	1197	1740		254	952	534	1740
	5.9%	25.3%	68.8%			14.6%	54.7%	30.7%	
	2111	1329	1780	5220		1592	2332	1296	5220
				45.79%					45.92%

Table A-8. Subject 16 Confusion Matrices.

Individually Trained ANN **Internal Validation** individual features	2636	214	9	2859	**Subject 16**
	92.2%	7.5%	0.3%		
	128	2700	119	2947	
	4.3%	91.6%	4.0%		Confusion Matrix Key
	5	62	2827	2894	
	0.2%	2.1%	97.7%		
	2769	2976	2955	8700	
				93.83%	
Individually Trained ANN **External Validation** individual features	4574	1225	1	5800	
	78.9%	21.1%	0.0%		
	1506	4029	265	5800	
	26.0%	69.5%	4.6%		
	13	294	5493	5800	
	0.2%	5.1%	94.7%		
	6093	5548	5759	17400	
				81.01%	

low CA	low C'd as med	low C'd as OV	TRUE low
med C'd as low	**medium CA**	med C'd as OV	TRUE medium
OV C'd as low	OV C'd as med	**overload CA**	TRUE overload
Classified low	Classified medium	Classified overload	**Overall CA**

Table A-8 Continued. Subject 16 Confusion Matrices.

Individually	678	201	10	889	Group	623	193	20	836
Trained ANN	76.3%	22.6%	1.1%		Trained ANN	74.5%	23.1%	2.4%	
Internal	154	644	41	839	**Internal**	184	617	80	881
Validation	18.4%	76.8%	4.9%		**Validation**	20.9%	70.0%	9.1%	
2-features	0	73	809	882	2-features	17	70	806	893
	0.0%	8.3%	91.7%			1.9%	7.8%	90.3%	
	832	918	860	2610		824	880	906	2610
				81.65%					78.39%
Individually	1030	709	1	1740	Group	1112	620	8	1740
Trained ANN	59.2%	40.7%	0.1%		Trained ANN	63.9%	35.6%	0.5%	
External	790	854	96	1740	**External**	811	782	147	1740
Validation	45.4%	49.1%	5.5%		**Validation**	46.6%	44.9%	8.4%	
2-features	0	36	1704	1740	2-features	22	37	1681	1740
	0.0%	2.1%	97.9%			1.3%	2.1%	96.6%	
	1820	1599	1801	5220		1945	1439	1836	5220
				68.74%					68.49%
Individually	769	123	3	895	Group	663	173	30	866
Trained ANN	85.9%	13.7%	0.3%		Trained ANN	76.6%	20.0%	3.5%	
Internal	108	706	34	848	**Internal**	222	612	53	887
Validation	12.7%	83.3%	4.0%		**Validation**	25.0%	69.0%	6.0%	
5-features	0	21	846	867	5-features	18	34	805	857
	0.0%	2.4%	97.6%			2.1%	4.0%	93.9%	
	877	850	883	2610		903	819	888	2610
				88.93%					79.69%
Individually	1293	447	0	1740	Group	1217	493	30	1740
Trained ANN	74.3%	25.7%	0.0%		Trained ANN	69.9%	28.3%	1.7%	
External	415	1224	101	1740	**External**	637	951	152	1740
Validation	23.9%	70.3%	5.8%		**Validation**	36.6%	54.7%	8.7%	
5-features	0	1	1739	1740	5-features	0	28	1712	1740
	0.0%	0.1%	99.9%			0.0%	1.6%	98.4%	
	1708	1672	1840	5220		1854	1472	1894	5220
				81.53%					74.33%
Individually	718	104	3	825	Group	736	123	7	866
Trained ANN	87.0%	12.6%	0.4%		Trained ANN	85.0%	14.2%	0.8%	
Internal	111	730	35	876	**Internal**	214	601	47	862
Validation	12.7%	83.3%	4.0%		**Vaidation**	24.8%	69.7%	5.5%	
10-features	0	22	887	909	10-features	42	45	795	882
	0.0%	2.4%	97.6%			4.8%	5.1%	90.1%	
	829	856	925	2610		992	769	849	2610
				89.46%					81.69%
Individually	1444	295	1	1740	Group	1529	197	14	1740
Trained ANN	83.0%	17.0%	0.1%		Trained ANN	87.9%	11.3%	0.8%	
External	447	1197	96	1740	**External**	569	1034	137	1740
Validation	25.7%	68.8%	5.5%		**Validation**	32.7%	59.4%	7.9%	
10-features	0	16	1724	1740	10-features	16	81	1643	1740
	0.0%	0.9%	99.1%			0.9%	4.7%	94.4%	
	1891	1508	1821	5220		2114	1312	1794	5220
				83.62%					80.57%
Individually	747	102	4	853	Group	766	82	16	864
Trained ANN	87.6%	12.0%	0.5%		Trained ANN	88.7%	9.5%	1.9%	
Internal	104	753	31	888	**Internal**	194	648	55	897
Validation	11.7%	84.8%	3.5%		**Validation**	21.6%	72.2%	6.1%	
15-features	1	32	836	869	15-features	11	36	802	849
	0.1%	3.7%	96.2%			1.3%	4.2%	94.5%	
	852	887	871	2610		971	766	873	2610
				89.50%					84.90%
Individually	1495	243	2	1740	Group	1564	171	5	1740
Trained ANN	85.9%	14.0%	0.1%		Trained ANN	89.9%	9.8%	0.3%	
External	410	1217	113	1740	**External**	481	1119	140	1740
Validation	23.6%	69.9%	6.5%		**Validation**	27.6%	64.3%	8.0%	
15-features	0	15	1725	1740	15-features	2	33	1705	1740
	0.0%	0.9%	99.1%			0.1%	1.9%	98.0%	
	1905	1475	1840	5220		2047	1323	1850	5220
				85.00%					84.06%

Bibliography

1. Air Force Research Laboratory | AFRL, (1998) "Flight Psychophysiology Laboratory," Office Brochure, Flight Psychophysiology Laboratory, Human Interface Technology Branch, Crew System Interface Division, Human Effectiveness Directorate (AFRL/HECP)

2. Auten, J. "G-LOC: Is the Cluebag Half Full or Half Empty?" *Flying Safety*, Vol 52, June 1996, pp 5-6.

3. Bauer, K. W. *OPER685, Applied Multivariate Data Analysis, Fall 1998*, Class Notes, Air Force Institute of Technology, OH.

4. Belue, L.M. (1992) *An Investigation of Multilayer Perceptrons for Classification*, M.S. Thesis, Air Force Institute of Technology, OH.

5. Belue, L.M. and Bauer, K.W. (1995) "Determining Input Features for Multilayer Perceptrons," *Neurocomputing*, Vol 7, pp.111-121.

6. Bishop, Christopher M. (1995) *Neural Networks for Pattern Recognition* (Oxford University Press Inc., New York, NY).

7. Bose, N. K. and P. Liang (1996) *Neural Network Fundamentals with Graphs, Algorithms, and Applications*. (McGraw-Hill, Inc., New York, NY).

8. Brookings, J. B., Wilson, G. F., Swain, C.R.(1996) "Psychophysiological Responses to Changes in Workload During Simulated Air Traffic Control," Biological Psychology, Vol 42, pp.361-378.

9. Burden, Richard L. and Faires, J. Douglas (1997) *Numerical Analysis*, Sixth Edition (Brooks/Cole Publishing Company, Pacific Grove, CA), pp. 526-537.

10. Caldwell, J.A., Roberts, K.A., Kelly, C. F., Jones, H.D., Lewis, J.A., Woodrum, L, Dillard, R.M., and Johnson, P.P. (1997) *Effects of Pilot Workload on EEG Activity Recorded During the Performance of In-Flight Maneuvers in a UH-1 Helicopter*, USAARL Report No. 97-31, U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL.

11. Comstock, J.R. and Arnegard, R.J., (1992) *The Multi-Attribute Task Battery for Human Operator and Strategic Behavior Research*, NASA Technical Memorandum 104174.

12. Defense Advanced Research Projects Agency (1988) *DARPA Neural Network Study*, AFCEA International Press, Fairfax, VA.

13. Demuth, H. and M. Beale (1998). *MATLAB Neural Network Toolbox User's Guide* (MathWorks, Natick, MA).

14. Dillon, W.R. and M. Goldstein. *Multivariate Analysis: Methods and Applications*, New York, NY: John Wiley & Sons, 1984.

15. Galley, N. (1993) "The Evaluation of the Electooculogram as a Physiological Measuring Instrument in the Driver Study of Driver Behaviour," *Ergonomics*, Vol 36, No. 9, pp. 1063-1070.

16. Gevins, A.S. and Leong, H.M.F. (1992) *Physiological Indices of Mental Workload*, Interim Technical Report, AFSOR Contract F49620-92-C-0013, 15 Dec 91 to 14 Dec 92, RN AD-A261 692.

17. Gevins, A.S., Smith, M.E., Leong, H., McEvoy, L., Whitfield, S., Du, R., and Rush, G. (1998) "Monitoring Working Memory Load during Computer-Based Tasks with EEG Pattern recognition Methods," *The Journal of the Human Factors and Ergonomics Society*, Vol 40, pp. 79-91.

18. Gevins, A.S., Zeitlin, G.M., Doyle, J.C., Yingling, C.D., Schaffer, R.E., Callaway, E., Yeager, C.L., (1979) "Electroencephalogram Correlates of Higher Cortical Functions," Science, Vol 203, pp. 665-668.

19. Greene, K.A. (1998) *Feature Saliency in Artificial Neural Networks with Application to Modeling Workload*, Ph.D. Dissertation, Air Force Institute of Technology, OH.

20. Greene, K.A., Bauer, K.W., Kabrinsky, M., Rogers, S.K., Russell, C.A., and Wilson, G.F.(1996) "A Preliminary Investigation of Selection of EEG and Psychophysiological Features for Classifying Pilot Workload," *Intelligent Engineering Systems through Artificial Neural Networks*, Vol 6, Dagli, C.H., Akay, M., Chen, C.L.P., Fernandez, B.R., and Ghosh, J. (eds), (ASME Press, New York, NY), pp. 691-697.

21. Greene, K.A., Bauer, K.W., Kabrinsky, M., Rogers, and Wilson, G.F.(1997) "Estimating Pilot Workload Using Elman Recurrent Neural Networks: A Preliminary Investigation," *Intelligent Engineering Systems through Artificial Neural Networks*,Vol 7, Dagli, C.H., Akay, M., Ersoy, O., Fernandez, B.R., and Smith A. (eds), (ASME Press, New York, NY), pp. 703-708.

22. Greene, K.A., Bauer, K.W., and Sumrell, D.B. "Feature Screening Using Signal-to-Noise Ratios," accepted by *Neurocomputing,* Aug 98.

23. Greene, K.A., Bauer, K.W., Wilson, G.F., Russell, C.A., Rogers, S.K., and Kabrinsky, M "Selection of Psychophysiological Features for Classifying Air Traffic Controller Workload in Neural Networks," submitted to *International Journal of Smart Engineering System Design.*

24. Griffiths, David J. (1989) *Introduction to Electrodynamics*, Second Edition (Prentice Hall, Englewood Cliffs, NJ), pp. 346-350.

25. Gundel, A. and Wilson, G.F. (1993) "Editorial – Psychophysiological Measures in Transport Operations," *Ergonomics*, Vol 36, No. 9, pp. 989.

26. Hanskins, T.C. and G.F. Wilson (1998) "A Comparison of Heart Rate, Eye Activity, EEG and Subjective Measures of Pilot Mental Workload During Flight," *Aviation, Space, and Environmental Medicine*, Vol 69, No. 4, pp. 360-367.

27. Jorna, P.G.A.M. (1993) "Heart rate and Workload Variations in Actual and Simulated Flight," *Ergonomics*, Vol 36, No. 9, pp. 1043-1054.

28. Levy-Leblond, Jean-Marc and Balibar, Francoise (1990) *Quantics – Rudiments of Quantum Physics* (Elsevier Science Publishing Company, Amsterdam, New York, NY) pp. 42-50.

29. Lizza, G.D. (1991) *Neural Network Classification of Mental Workload Conditions by Analysis of Spontaneous Electroencephalograms*, M.A. Thesis, Wright State University, OH.

30. MathWorks Inc., (1996) *MATLAB Signal Processing Toolbox User's Guide*, Version 4, (MathWorks, Natick, MA), pp. 2-2 – 2-45.

31. Morton, P.E. and Wilson, G.F. (1992) *Backpropagation and EEG Data*, Final Report for Period 19 Dec 88 to 30 Jun 89, Armstrong Aerospace Medical Research laboratory, Human Engineering Division, Air Force Systems Command, Wright-Patterson Air Force Base, OH, RN AD-A279 073.

32. Nelson, Marilyn McCord and Illingworth, W. T. (1991) *A Practical Guide to Neural Nets* (Addison-Wesley Publishing Company, Inc., NewYork, NY).

33. Quartz, S.R, Stensmo, M., Makeig, S., and Sejnowski, T.J. (1995) "Eye Blink Rate as a Practical Predictor for Vigilance," *Society for Neuroscience*, Vol 21, pp. 939.

34. Rizzo, C.W. (1998) *Parallel implementation of an Artificial Neural Network Integrated Feature and Architecture Selection Algorithm*, M.S. Thesis, Air Force Institute of Technology, OH.

35. Roscoe, A.H. (1993) "Heart Rate as a Psychophysiological Measure for In-Flight Workload Assessment," *Ergonomics*, Vol 36, No. 9, pp. 1055-1062.

36. Ruck, D.W. (1990) *Characterization of Multilayer Perceptrons and their Application to Multisensor Automatic Target Detection*, Ph.D. Dissertation, Air Force Institute of Technology, OH.

37. Ruck, D.W, Rogers, S.K., and Kabrisky, M. (1990) "Feature Selection Using a Multilayer Perceptron," *Journal of Neural Network Computing*, Vol 2, pp. 40-48.

38. Russell, C.A., Wilson, G.F., and Monett, C.T. "Mental Workload Classification Using a Backpropagation Neural Network," *Intelligent Engineering Systems through Artificial Neural Networks*, Proceedings of Artificial Neural Networks in Engineering (ANNIE) International Conference, St. Louis, MO, 10-13 Nov 1996, eds. C.H. Dagli, M. Akay, C.L.P. Chen, B.R. Fernandez, and J. Ghosh, Vol 6, pp685-690.

39. Russell, C.A. and Wilson, G.F. "Air Traffic Controller Functional State Classification Using Neural Networks," *Intelligent Engineering Systems through Artificial Neural Networks,* Proceedings of Artificial Neural Networks in Engineering (ANNIE) International Conference, St. Louis, MO, 1-4 Nov 1998, eds. C.H. Dagli, M. Akay, A.L. Buczak, O. Ersoy, and B.R. Fernandez, Vol 8, pp649-654.

40. Sirevaag, E.J., Kramer, A.F., Wickens, M.R., Strayer, D.L., and Grennel, J.F.(1993) "Assessment of Pilot Performance and Mental Workload in Rotary Wing Aircraft," *Ergonomics*, Vol 36, No. 9, pp. 1121-1140.

41. Smith, Murray. (1996) *Neural Networks for Statistical Modeling* (Thomson Publishing Inc., Boston, MA)

42. Steppe, J.M. (1994) *Feature and Model Selection in Feedforward Neural Networks*, Ph.D. Dissertation, Air Force Institute of Technology, OH.

43. Steppe, J.M. and Bauer, K.W. (1996) "Improved Feature Screening in Feedforward Neural Networks," *Neurocomputing*, Vol 13, pp.47-58.

44. Stern, J.A. and Dunham, D.N. (1990) *Principles of Psychophysiology: Physical, Social, and Inferential Elements,* (Cambridge University Press), pp. 513-553.

45. Sumrell, D.B. (1996) *An Investigation of Preliminary Feature Screening Using Signal-to-Noise Ratios* , M.S. Thesis, Air Force Institute of Technology, OH.

46. Tarr, G.L. (1991) *Multi-Layered Feedforward Neural Networks for Image Segmentation*, Ph.D. Dissertation, Air Force Institute of Technology, OH.

47. Wackerly, D.D., Mendenhall, W., Scheafer R.L. (1996) *Mathematical Statistics with Applications* (Duxbury Press, Belmont, CA).

48. Wasserman, Phillip P. (1989) *Neural Computing Theory and Practice* (Van Nostrand Reinhold, New York, NY).

49. Wiggins, V.L., Borden, K.M., Turner, K.L., Looper, L.T., and Grobman, J.H. (1996) *Statistical Neural Network Analysis Package (SNNAP) Version 2.0 User's Manual,*

Interim Technical Paper – July 1994 – July 1995, Air Force Material Command, Brooks Air Force Base, TX.

50. Wilson, G.F. (1993) "Air-to-Ground Training Missions: A Psychophysiological Workload Analysis," *Ergonomics*, Vol 36, No. 9, pp. 1071-1087.

51. Wilson, G.F. and F. T. Eggemeier (1991) "Psychological Assessment of Workload in Multi-Task Environments," *Multiple-Task Performance*, Damos, D.L.,(ed), (Taylor & Francis, London, UK), pp. 229-260.

52. Wilson, G.F. and F. Fisher (1991) "The Use of Cardiac and Eye Blink Measures to Determine Flight Segment in F4 Crews," *Aviation, Space, and Environmental Medicine,* Oct. 1991, pp. 959-962.

53. Wythoff, B.J., (1993) "Backpropagation Neural Networks -- A Tutorial," Chemometrics and Intelligent Laboratory Systems, Vol 18, pp. 115-155.

Vita

Captain Trevor I. Laine was born on 20 October 1970 in San Diego, California. After a couple years, his father parted from military service and the family moved back to Oregon. In 1988, he graduated from Woodrow Wilson High School in Portland, Oregon. After receiving an AFROTC scholarship, he then moved across the river and attended the University of Portland. In 1992, he completed his Bachelor of Science degree in Physics and received a commission in the USAF. His first assignment was to the Space and Missile Systems Center (SMC), Los Angeles AFB, California where he was assigned to the Titan IV System Program Office. After spending three years as the project officer in charge of the Centaur upper-stage liquid rocket engines, Trevor went on to spend his last year as a project officer in SMC's strategic planning office. Amidst his four-year tour and zest for beach life, he also completed his Master's in Business Administration from Chapman University. Trevor then entered the Air Force Institute of Technology in August of 1997. After graduation from the Air Force Institute of Technology, Trevor will be assigned as an analyst for the Office of Aerospace Studies at Kirtland AFB, New Mexico. Trevor's hobbies include numerous outdoor sports and activities with golf, snow skiing, and hunting & fishing to name a few.

Permanent Address: c/o John Laine
1521 S.W. Hume Ct.
Portland, OR 97219

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1999	Master's Thesis

4. TITLE AND SUBTITLE
SELECTION OF PSYCHOPHYSIOLOGICAL FEATURES ACROSS SUBJECTS FOR CLASSIFYING WORKLOAD USING ARTIFICIAL NEURAL NETWORKS

5. FUNDING NUMBERS

6. AUTHOR(S)
Trevor I. Laine, Captain, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology (AFIT)
2950 P Street
Wright-Patterson AFB, OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/GOR/ENS/99M-09

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Dr. John F. Tangney
AFOSR/NL
110 Duncan Ave, Suite B115
Bolling AFB, DC 20322-0001
(202) 767-8075 (DSN 297)

Dr. Glenn F. Wilson
AFRL/HECP
2255 H Street
Wright-Patterson AFB, OH 45433-7022
(937) 255-8748 (DSN 785)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Advisor was Dr. Kenneth W. Bauer, Jr., of AFIT/ENS (937) 255-6565 x4328 (DSN 785)
kenneth.bauer@afit.af.mil

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
The issue of pilot workload is important to the United States Air Force because pilot overload or task saturation leads to decreases in mission effectiveness. Additionally, in the most extreme cases, pilot overload may lead to the loss of aircraft and crewmember lives. Current research efforts are utilizing psychophysiological data including electroencephalography (EEG), cardiac, eye-blink, and respiration measures in an attempt to identify workload levels. The primary focus of this effort is to determine if a single parsimonious set of psychophysiological features exists for accurately classifying workload levels between multiple test subjects. To accomplish this objective, the signal-to-noise (SNR) saliency measure is used to determine the usefulness of psychophysiological features in feedforward artificial neural networks (ANN). The SNR saliency measure determines the saliency, or relative value, of a feature by comparing it to a feature of injected noise. For this effort, 36 psychophysiological features were derived from the data collected as each subject completed simulated crewmember tasks using the Multi-Attribute Task Battery developed by NASA. These tasks were randomly presented to the subjects in blocks with three distinct levels: low, medium, and an overload level in which subjects could not complete all tasks.

14. SUBJECT TERMS
Feature saliency, Feature selection, Neural networks, Discriminant analysis, Workload

15. NUMBER OF PAGES
163

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL